DOCUMENT RESUME

ED 309 952 SE 050 788

AUTHOR Hedges, Larry V.; And Others
TITLE A Practical Guide to Modern Methods of Meta-Analysis.
INSTITUTION National Science Teachers Association, Washington, D.C.
SPONS AGENCY National Science Foundation, Washington, D.C.
REPORT NO ISBN-0-87355-081-1
PUB DATE 89
GRANT NSF-MDR-8550470
NOTE 80p.
AVAILABLE FROM National Science Teachers Association, 1742 Connecticut Avenue, NW, Washington, DC 20009 ($9.50; PB-52).
PUB TYPE Guides - Classroom Use - Materials (For Learner) (051) -- Books (010) -- Reports - Research/Technical (143)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Comparative Analysis; Effect Size; Higher Education; *Meta Analysis; *Research Methodology; *Science Education; *Statistical Analysis; Statistical Data; Statistical Studies; Synthesis

ABSTRACT

Methods for meta-analysis have evolved dramatically since Gene Glass first proposed the term in 1976. Since that time statistical and nonstatistical aspects of methodology for meta-analysis have been developing at a steady pace. This guide is an attempt to provide a practical introduction to rigorous procedures in the meta-analysis of social science research. It approaches the use of modern statistical methods in meta-analysis from the perspective of a potential user. The treatment is limited to meta-analysis of studies of between-group comparisons using the standardized mean difference as an index of effect magnitude. This guide is organized according to a variant of Cooper's stages of the research review process: (1) problem formulation; (2) data collection and data evaluation, data analysis and interpretation; and (3) presentation of results. Although each stage is discussed, the greatest emphasis is placed on the stage of data analysis and interpretation. Examples from a synthesis of research on the effects of science curricula are used throughout for illustration. Because this book is intended to be a practical guide, the references are provided primarily to exemplify issues or techniques rather than to provide theoretical discussions or derivations. (CW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED309952

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
Michael Shackelford

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

SE 050 788

A PRACTICAL GUIDE TO MODERN METHODS OF META-ANALYSIS

Larry V. Hedges, James A. Shymansky & George Woodworth

Copyright © 1989 by the National Science Teachers Association, 1742 Connecticut Avenue NW, Washington, DC 20009. All rights reserved. This volume may not be reproduced in whole or in part in any form without written permission from the National Science Teachers Association.

Produced by Special Publications
National Science Teachers Association
1742 Connecticut Avenue NW
Washington, DC 20009

Shirley Watt, *Managing Editor*
Peter Andersen, *Editorial Assistant*
Elizabeth McNeil, *Editorial Assistant*

Stock Number PB 52
ISBN 0-87355-081-1

This guide was prepared with support from the National Science Foundation under Grant No. MDR-8550470. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the National Science Foundation.

CONTENTS

Introduction		iii
Preface		v
Meta-analysis as a form of research		1
1.0	**Problem Formulation**	2
1.1	Confirmatory versus exploratory reviews	2
1.2	Deciding which studies can be aggregated	3
1.3	Selecting constructs and operations	4
1.4	Broad versus narrow operations for constructs	5
2.0	**Data Collection**	6
2.1	Sampling in meta-analysis	6
2.1.1	Identifying a target population	7
2.1.2	Searching for possible relevant studies	8
2.2	Missing data in meta-analysis	9
2.2.1	Missing data on study outcome	9
2.2.2	Missing data on study characteristics	10
2.3	Publication bias	11
2.4	Establishing coding procedures	11
2.4.1	Identifying contrasts within studies	12
2.4.1.1	Study ID	12
2.4.1.2	Author/reference	12
2.4.1.3	Sample ID	12
2.4.1.4	Dependent variable	13
2.4.1.5	Time of testing	13
2.4.2	Study context	13
2.4.3	Subject characteristics	13
2.4.4	Study design and execution	14
2.4.4.1	Controls for pre-existing difference	14
2.4.4.2	Experimental mortality	15
2.4.4.3	Treatment contamination of control groups	15
2.4.4.4	Unit of analysis	15
2.4.5	Treatment	16
2.4.5.1	Length of treatment	16
2.4.5.2	Treatment fidelity	16
2.4.6	Control groups	17
2.4.7	Outcome variables	17
2.5	Organizing data: Designing a coding sheet	18
2.5.1	Sample coding sheet	20

2.5.2	Coding protocols and data screening	20
2.6	Reliability of coding	21
3.0	**Data Analysis and Interpretation**	23
3.1	Effect sizes and effect size estimates	23
3.1.1	Effect size estimates	23
3.1.2	Correction for bias	25
3.2	Estimating effect size when means and standard deviations are not reported	26
3.2.1	Calculating effect size estimates when there are means and ANOVA tables, but no standard deviations	26
3.2.2	Calculating effect size estimates when only t or F statistics are reported	30
3.2.2.1	Exact effect sizes from t or single-factor F	30
3.2.2.2	Approximate effect sizes from multifactor ANOVA or ANCOVA	30
3.2.3	Sometimes effect sizes cannot be estimated	33
3.2.3.1	Paired designs	33
3.2.3.2	Insignificant differences	34
3.3	Standard error of effect size estimates	34
3.4	Combining effect size estimates	36
3.4.1	Combining independent effect size estimates by weighted averaging	36
3.4.2	Combining correlated effect size estimates	37
3.5	When is it appropriate to combine estimates: Measuring heterogeneity	39
3.5.1	Heterogeneity plots	42
3.6	Formal analysis of heterogeneity: An analysis of variance for effect sizes	43
3.7	Combining effect size estimates with unexplained heterogeneity: Random effects models in meta-analysis	43
3.7.1	Estimating the variance component	44
3.7.2	Combining effect size estimates in random effects models	45
4.0	**Reporting Results**	47
References		50
Appendix I	Revised meta-analysis coding sheet	53
Appendix II	Technical notes: Pooling correlated effect sizes: Notation	56
Appendix III	Computer SAS programs	57
Appendix IV	Achievement effect sizes and heterogeneity plots from Shymansky, Hedges, and Woodworth	63

INTRODUCTION

In June of 1984, I was asked by the NSTA editorial staff to examine some of the material on meta-analysis that would be included in one of the National Science Teachers Association's books in the *What Research Says to the Science Teacher* series. Because of my background in physics, for which I had received considerable mathematical training in statistics and error theory, our editor thought I could help assess the accuracy of the material that we were to publish.

Although I have admittedly little knowledge or background in educational statistics, what I read in the galley proofs of the new monograph bothered me. Quantities (in this case meta-analysis effect sizes) determined from independent studies, and therefore, independent measurements, were being averaged, without taking into account their relative precisions. In the theory of errors, as used in the sciences, it is well-known that independent measurements can be averaged properly only if each measurement is weighted by the inverse of its variance of error. In this way, the most precise measurement contributes the greatest amount to the mean which results.

Because weighting had not been done for the material to be published in NSTA's monograph, I halted publication, and advised the authors, asking that a reanalysis be conducted before NSTA would publish the material. I offered to help secure NSF funding to support the effort. I was very much concerned that policy-makers not use conclusions from a faulty meta-analysis to make important decisions in science education, and certainly not from anything NSTA had published.

In the process of considering reanalysis, James Shymansky, one of the authors of this handbook, found recent articles on the subject in the literature; several were written by Larry Hedges, one of the other authors of the handbook. I read several of the articles and communicated directly with Hedges. It was clear that Hedges had done important theoretical work on proper weighting of effect sizes, and he had done so well before I observed the problem in the material NSTA was to publish. In fairness to the researchers involved in the material NSTA rejected, it should be pointed out that their research design and data collection had been initiated well before most of Hedges' brilliant work on the topic had been published. The techniques for weighting effect sizes are presented systematically and clearly in section 3.0 of this handbook. With NSF support, Shymansky, Hedges, and George Woodworth, a statistician at Iowa, carried out a reanalysis of the data in question, and they have separately published those results. As part of their NSF project, they produced this excellent handbook for the research practitioner, so that the power of meta-analysis can be utilized by the educational or social science researcher, without having to try to understand the complexities of its mathematical foundations.

Having tried to read the various theoretical papers on meta-analysis, I am particularly impressed with the clarity with which this handbook addresses applications of meta-

analysis to the kinds of research questions common to science education. The National Science Teachers Association is pleased to make this excellent publication available to the community of researchers.

Bill G. Aldridge
Executive Director
National Science Teachers Association

PREFACE

This guide is designed to be a practical introduction to the application of modern methodology for meta-analysis. Methods for meta-analysis have evolved dramatically since Gene Glass first proposed the term in 1976. Since that time statistical and nonstatistical aspects of methodology for meta-analysis have been developing at a steady pace. One very important methodological development has been the flowering of work on statistical methods designed particularly for meta-analysis. There are now six books and well over 100 articles that treat methods for meta-analysis in the social sciences. With all this literature currently available, the addition of yet another monograph requires some justification.

The reason for this guide is that the existing literature on methods for meta-analysis, although voluminous, is deficient. The journal articles are widely scattered in many journals (although the collections by Rosenthal, 1980 and 1984; Light, 1983; and Yeaton and Wortman, 1984 are highly recommended). The existing books do an admirable job of bringing together the diverse literature but their treatment of statistical methods is uneven. The book by Glass, McGaw, and Smith (1981), for example, is a classic work, but it was written before the development of most of the modern statistical methodology that has become the state of the art for meta-analysis. The book by Hunter, Schmidt, and Jackson (1982) provides an excellent introduction to the random effects models developed for the study of validity generalization and their application to meta-analyses involving standardized mean differences. It does not, however, provide a treatment of the very considerable literature on fixed and mixed effects models that have developed since the book was written. The book by Light and Pillemer (1984) provides an outstanding introduction to the conceptual aspects of meta-analysis, but it does not treat specific statistical methods in detail. The short book by Cooper (1984) is a lucid introduction to procedures for rigorous research syntheses, but it also provides only a brief introduction to statistical methodology for meta-analysis. The book by Rosenthal (1984) provides an extraordinarily clear description of the statistical methods treated, but it does not treat many of the most powerful and widely used statistical methods for meta-analysis. Finally, the book by Hedges and Olkin (1985) treats statistical methods in exhaustive detail, but does so at a technical level that is inaccessible to some social scientists.

This guide is an attempt to provide a practical introduction to rigorous procedure in the meta-analysis of social science research. It approaches the use of modern statistical methods in meta-analysis from the perspective of a potential user. The treatment is limited to meta-analysis of studies of between-group comparisons using the standardized mean difference as an index of effect magnitude. It does not address the meta-analysis of correlation coefficients.

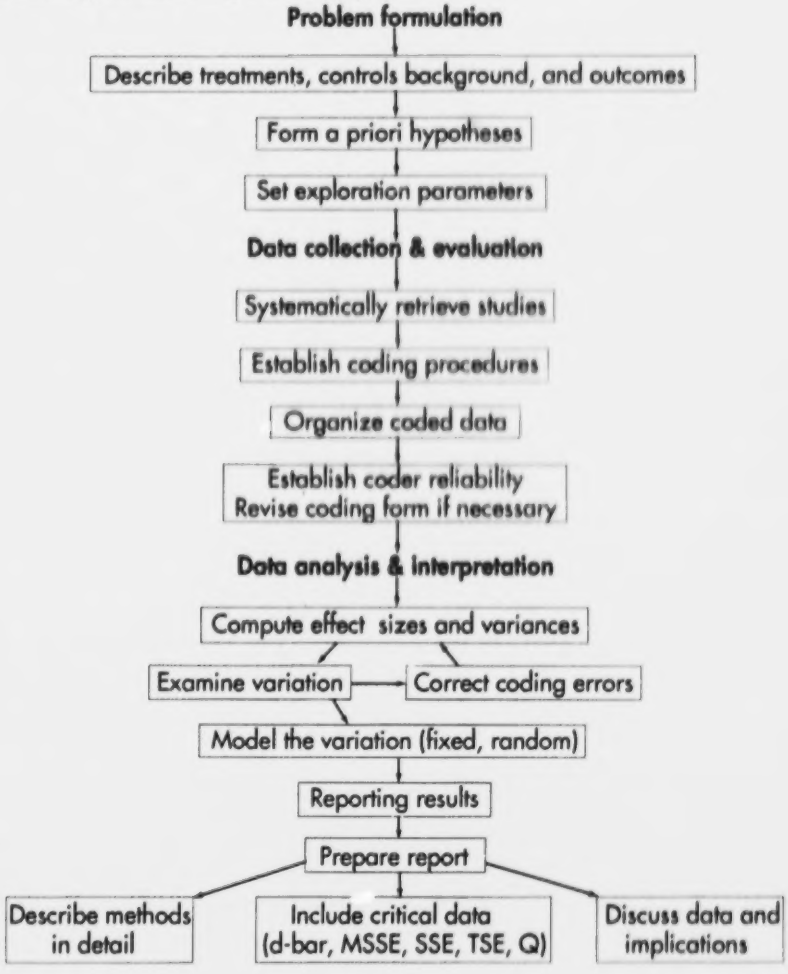
This guide is organized according to a variant of Cooper's stages of the research

review process: (1) problem formulation, (2) data collection and data evaluation, data analysis and interpretation, and (3) presentation of results. Although each stage is discussed, the greatest emphasis is placed on the stage of data analysis and interpretation. Examples from a synthesis of research on the effects of science curricula are used throughout for illustration. Because this is intended to be a practical guide, the references are provided primarily to exemplify issues or techniques rather than to provide theoretical discussions or derivations. Complete bibliographies of theoretical references on qualitative aspects of methodology are in Cooper (1984), Light and Pillemer (1984), and Rosenthal (1984). Hedges and Olkin (1985) provide a complete bibliography on statistical methods for the meta-analysis of standardized mean differences.

Meta-analysis as a form of research

Original research proceeds through a series of stages which distinguishes it as "disciplined inquiry" (Shulman, 1986). The inquiry process begins with a problem formulation, moves to a stage of data collection and evaluation, is followed by data analysis and interpretation, and culminates in a final report. So it is with meta-analysis; it too proceeds through similar stages of inquiry. Discussions and examples in the guide will follow the steps illustrated in Figure 1.

FIGURE 1: Steps in conducting a meta-analysis



1.0 Problem formulation

1.1 Confirmatory versus exploratory reviews

Problem formulation is the first step in any research study or research review. It involves formulating the precise questions to be answered. One aspect of formulating questions is deciding whether your review is to be confirmatory (hypothesis testing) or exploratory (hypothesis generating). Obviously, new hypotheses (even new variables) arise in the course of meta-analyses, just as in any scientific activity. But you must distinguish the clearly *a priori* hypotheses from those that are suggested by the data. This distinction has implications for the choice of statistical analysis procedures used in the meta-analysis. Techniques described later in this guide are designed for reviews whose principal mode of inquiry is to test hypotheses that are formed *a priori*. Using these procedures when hypotheses are not formed in advance can be misleading. Similarly, when many statistical analyses are conducted on the same data, the usual significance levels will *not* reflect the chance of making *at least* one Type I error in the collection of tests (the simultaneous significance level). Thus, when conducting many tests in an "exploratory" mode, there is a tendency to "capitalize on chance."

The problems of exploratory analysis can be dealt with in several ways. One way is to use statistical methods that are specifically designed for exploratory analysis such as clustering methods (Hedges & Olkin, 1983, 1985). Alternatively, you may adjust the significance level when many tests are conducted on the same data. The problem with this and all other simultaneous procedures is that they reduce the power of the statistical tests and the effect is dramatic when many tests are conducted simultaneously.

You may also use procedures that do not involve statistical significance. Descriptive procedures are the simplest to use. Graphical procedures such as Light and Pillemer's (1984) funnel diagrams, Hedges and Olkin's (1985) confidence interval plots, or many of the graphical ideas presented by Tukey (1977) may also be used. In a later section of this guide procedures for determining confidence interval plots are presented.

If your data set is large enough, you might also consider dividing your data into two subsets. You can use one subset to generate hypotheses whose statistical significance can then be evaluated (cross-validated) on the second subset (Light & Pillemer, 1984).

Example 1: Problem formulation

> In the meta-analysis of the effects of new science curricula by Shymansky, Hedges, and Woodworth (1986), a principal question to be answered was whether the new process oriented science curricula produced higher achievement and more positive attitudes than traditional (pre 1960's) science curricula. This question was formulated a priori and the procedures used in the meta-analysis were directed toward testing the hypothesis that new science curricula produced greater achievement and more positive attitudes.
>
> In this same meta-analysis, effects were found to be inconsistent across studies. In

searching for the factors associated with the variability among effects, a variety of possible hypotheses were constructed to explain the variability. For example, we developed the hypothesis that the type of control group used and the degree of pre-existing difference between groups was related to the variability in effect sizes.

1.2 Deciding which studies can be aggregated

A second aspect of problem formulation concerns decisions about when studies are examining "the same problem." That is, you must decide whether treatments, controls, experimental procedures, and study outcome measures are comparable enough to be considered the same for the purposes of aggregation. Although different studies examining the same problem will not be identical, it is often possible to distinguish studies that conceptualize the treatment, controls, or outcome in the same way. It is helpful to distinguish the *theoretical* or *conceptual* variables about which knowledge is sought (called "constructs") from the actual *examples* of these variables that appear in the studies (called "operations").

Example 2: Identifying constructs and operations

Suppose we want to know if a particular method of teaching mathematics leads to better problem solving. To find out, a comparative study is conducted in which students are randomly assigned to teachers, some of whom use the new method. A problem solving test is then administered to students to determine which group of students were better at problem solving. The exact conceptualization of mathematical problem solving is a construct. The particular test used to measure problem solving is an operation corresponding to that construct.

Similarly, a particular teaching method as defined conceptually is a construct (e.g., inquiry teaching), while the behavior of a particular teacher trying to implement that teaching method is an operation (e.g., questioning). The point here is that even when studies share the same constructs, they almost surely differ in the operations that correspond to those constructs.

Defining questions precisely in a research review involves deciding on the constructs of independent variables, study characteristics, and outcomes that are appropriate for the questions addressed by the review and deciding on the operations that will be regarded as the corresponding constructs. That is, you must develop both the construct definitions and a set of rules for deciding which concrete instances of treatments, controls, or measures correspond to those constructs.

Although the questions of interest might seem completely self-evident in a review, a little reflection may convince you that there are subtleties in formulating precise questions.

Example 3: Problems in defining the question

Consider the seemingly well-defined question, "Are discovery learning science curricula superior to conventional science instruction?" What problems does this question pose? Here are a few:

- What exactly is meant by the term "discovery learning"?
- Does discovery learning imply activity at the single student or small group level or should group-based discovery-oriented activities also count?
- Should curricula intended as discovery but modified by teachers to be more "teacher-centered" be considered discovery or conventional?
- What exactly is meant by the term "conventional science instruction"?

Before proceeding with your meta-analysis, you should think carefully about the problem under review. Have you clearly defined the construct definitions and established a set of rules for deciding which instances of treatments, controls, and outcomes you are going to include in your review? Time spent specifying these parameters at the outset of your study will be time saved in the later stages of data collection and analysis.

1.3 Selecting constructs and operations

One of the potential problems of meta-analysis (or any research synthesis) is that they may combine incommensurable evidence (sometimes referred to as the case of "apples and oranges"). This is essentially a criticism of the breadth of constructs and operations chosen. In one sense the breadth of constructs and operations chosen must reflect the breadth of the question addressed by the review. Constructs and operations used in a review should usually be distinguished more narrowly by the reviewer than may be reflected in the final presentation of results. Thus, the issue is first what constructs and operations are to be *included* in the review, then what constructs and operations are to be *distinguished* in the data analysis of the review, and finally which constructs and operations are to be *presented* in the results of the review.

Meta-analysts have tended to use rather broad constructs and operations in their presentation of results. This may be due to the ease with which quantitative methods can analyze data from large numbers of studies. It is important to recognize, however, that while broad questions necessitate the inclusion of studies with a broad range of constructs and operations, they need not inhibit the meta-analyst from distinguishing variations of these constructs in the data analysis and in presentation of results.

Perhaps the most successful applications of broad or multiple constructs in meta-analysis are those that *include* broad constructs in the data analysis and presentation. This approach permits you to examine variations in the pattern of results as a function of construct definition. It also permits you to analyze separately each narrow construct (see e.g., Cooper, 1979; Linn & Peterson, 1985; Eagly & Carli, 1981; Thomas & French, 1985). You may carry out a combined analysis across constructs where appropriate or present distinct analyses for the separate constructs.

Example 4: Focusing the problem statement

> In the study of the effects of new science curricula on student performance (Shymansky, Hedges, & Woodworth, 1987), the broad constructs of "new science curricula" and "student performance" were further distinguished. For example, new science curricula were broken down by subject area, grade level, specific curriculum project, and student gender and outcome measures were broken into criterion groups such as achievement, attitudes, and process skills. Analyses were then performed on the specific subgroups and on crosses between these subgroups (e.g., student achievement for boys versus girls by subject area).

1.4 Broad versus narrow operations for constructs

Another issue arises at the level of operationalization of constructs. You will always have to admit several different operations for a given construct. Treatments will not have been implemented identically in all studies and different studies will not have measured the outcome constructs in exactly the same way. Thus, you will have to judge whether each operation is a legitimate representation of the corresponding construct. This means you will have to obtain as much information about the treatment actually implemented and the outcome measure actually used in each study. For this you may have to go to a secondary source such as technical reports, test reviews, or published tests.

Example 5: Clarifying the constructs and operations

> In the Shymansky, et al., study of new science curricula, some studies were encountered in which teachers had modified the new curricula to fit personal styles or school policy. Since this can be expected when teachers adopt a new program in most school situations, the decision was made to include the study. In another case, however, some studies were found in which "no science instruction" was used as the conventional science instruction comparison. This was not considered a fair comparison and these studies were excluded from the analysis.

In spite of the difficulty multiple operations may cause you in your review, they can also enhance the confidence in relationships if the analogous relationships between operations hold under a variety of different operations (Campbell, 1969). However, increased confidence comes from multiple operations only when the different operations are in fact more related to the construct under study than to some other construct (see the discussion of multiple operationalism in Webb, Campbell, Sechrest, & Grove, 1981). Thus, although multiple operations can lead to increased confidence through "triangulation" of evidence, the indiscriminate use of broad operations can also contribute to invalid results by confounding one construct with another (see Cooper, 1984).

2.0 Data Collection

Data collection in meta-analysis consists of assembling a body of research studies and extracting quantitative indices of study characteristics and of effect magnitude. The former is largely a problem of selecting studies that may contain information relevant to the specific questions you want to address. It is a sampling process. The latter is a problem of obtaining quantitative representations of the measures of effect magnitude and the other characteristics of studies that are relevant to the specific questions you want to address. It is a measurement process. The standard psychological measurement procedures for ensuring the reliability and validity of such ratings are as appropriate in meta-analysis as in original research. In this section of the guide we will discuss procedures for collecting and evaluating study data that have proven effective in previous meta-analyses.

2.1 Sampling in meta-analysis

The problem of assembling a collection of studies is often viewed as a sampling problem: the problem of obtaining a representative sample of all studies that have actually been conducted. Because the adequacy of your sample necessarily determines the range of valid generalizations that are possible, the procedures you use in locating the studies for your meta-analysis are crucial. Much of the discussion on sampling in meta-analysis concentrates on the problem of obtaining a representative or exhaustive sample of studies that have been conducted. But this is not the only or even the most important aspect of sampling in meta-analysis. The more important sampling question is whether the sample of subjects and treatments in the individual studies are representative of the subject and treatment populations of interest.

The importance of representative sampling of subjects is obvious.

Example 6: Problems of nonrepresentative sampling of subjects

Studies of the effects of psychotherapy on college students who do not have psychological problems might be considered *nonrepresentative* in a meta-analysis to determine the effects of psychotherapy on patients with real psychological problems.

But the importance of representative sampling of treatments is perhaps more subtle. The question is whether the treatments which occur in the study are representative of the treatments about which you are seeking knowledge.

Example 7: Problems of nonrepresentative sampling of treatments

Studies of individualized or self-paced instructional methods might be considered nonrepresentative in a meta-analysis to determine the effects of new science curricula on student performance simply because the methods don't necessarily stress laboratory-based, inquiry activity.

The problem of obtaining representative samples of subjects and treatments is constrained by the sampling of studies and consequently is not under your complete control. You can, however, present descriptions of the samples of subjects and treatments and examine the relationship between the sample descriptions and study outcomes.

2.1.1 Identifying a target population

The first step in developing a data collection (sampling) plan in a meta-analysis is to define the target population of studies. You might think that the definition should include *all* studies of a particular problem, but it is desirable in practice to limit the target population. For instance, you might want to limit studies to the use of a particular general category of study methodology or procedure.

Example 8: Targeting studies by methodologies

- laboratory studies (of sex differences in conformity)
- field studies (of nonverbal communication)
- randomized experiments (of the effects of desegregation)
- quasi-experimental studies (of peer tutoring programs)

Or you might target the population by specifying the time period in which the studies were conducted.

Example 9: Targeting studies by time period

- studies of sex differences in cognitive abilities published between 1960 and 1985
- studies of the effects of new science curricula conducted between 1958 and 1972

You might limit the target population by specifying the setting or type of subject.

Example 10: Targeting studies by subject type

- studies of the effects of discovery learning in public elementary schools
- studies of the effects of direct instruction in urban secondary schools
- studies of the effects of visually-based instruction with low SES middle school children
- studies of the effects of peer-tutoring with college students

Or you might limit the target population by specifying the particular variations of treatment, outcome, or controls.

Example 11: Targeting studies by variable

- studies of the effectiveness of behavioral therapy for phobias
- studies of the effects of contingency contracting on arrest rates in juvenile delin-

quency
- studies of the effects of intensive practice and coaching on verbal intelligence test scores
- studies of the effectiveness of introspective therapies versus a placebo

Another strategy for limiting study populations involves the use of the medium in which the study was published, such as specifying all relevant journal articles or all articles published in a particular journal. But this procedure is not without problems either because published studies or those published in a particular journal may not be representative of all studies actually conducted. Published studies may not be representative because studies yielding statistically insignificant results are less likely to be published (see section 2.3 for more discussion on publication bias and methods for dealing with it). Studies published in particular journals may not be representative of all studies because journals are often part of citation networks which tend to use similar methodologies or have similar theoretical predispositions.

2.1.2 Searching for possible relevant studies

Once you have defined the target population, a systematic search for possibly relevant studies can be undertaken. The principal tools for systematic literature searches are the abstracting and indexing services for the social sciences. The three most prominent services are the Education Resources Information Center (ERIC) system, Psychological Abstracts, and Dissertation Abstracts International (DAI). The ERIC system produces two monthly indexes, *Resources in Education*, which is a guide to non-published documents and *Current Index to Journals in Education* which is a guide to journal literature in education. Psychological Abstracts publishes a single monthly guide to journal articles in psychology. DAI focuses exclusively on doctoral dissertations and publishes a guide to recent dissertations.

A different type of indexing service that is often useful is the Social Science Citation Index. This index provides a listing of journal articles that cite a given article. It can be used, for example, to find all related articles that cite a seminal article in the area of interest.

These four indexing services are not the only ones available. Karl White's *Sources of Information in the Social Sciences* (1973) provides a broad overview of hundreds of abstracting and indexing services.

Computerized searches of abstracting and indexing services are useful for several reasons. First, a computer search is much quicker and more cost effective than a manual search. It is not unthinkable to do several searches using a slightly different set of key words. In fact, you should plan on doing several searches. The best search is usually produced after several attempts to lead to successive refinements of key words or descriptors with which you can home-in on the potentially relevant studies.

A second advantage of computerized literature searches is that they often permit searching of titles and abstracts for key words or phrases. By scanning the abstracts you are more likely to identify relevant studies than with a scan of the much more limited list

of key words provided by the author or the abstractor in cataloging the article.

A third advantage of computerized searches is that you can receive printed copies of abstracts of the studies identified. Working from a printed abstract, you can screen out studies that may not be relevant to your review. This will save you time tracking down the full text of articles which are irrelevant to your review.

2.2 Missing data in meta-analysis

Missing data is a problem that plagues many forms of applied research. Survey researchers are well aware that the best sampling design is ineffective if the information sought cannot be extracted from the units that are sampled. Of course, missing data is not a problem if it is "missing at random," that is, if the missing information is essentially a random sample of all the information available. But this is an assumption which, if not true, may pose a serious threat to the validity of conclusions in meta-analysis. The specific cases of missing data on study outcome and study characteristics are considered separately in the next two sections.

2.2.1 Missing data on study outcome

Studies (such as single case studies) that do not use statistical analyses are one source of missing data on study outcome. Of the studies that use statistics, some do not provide enough statistical information to allow the calculation of an estimate of the appropriate outcome parameter. Sometimes this is a consequence of failure to report relevant statistics. More often it is a consequence of the researcher's use of a complex design that makes difficult or impossible the construction of a parameter estimate that is completely comparable to those of other studies. Unfortunately, both the sparse reporting of statistics and the use of complex designs are plausibly related to study outcomes. Both result, at least in part, from the editorial policies of some journals which permit publication of only the most essential statistics. Perhaps the most pernicious sources of missing data are studies that *selectively* report statistical information. Such studies typically report only those results which are statistically significant, exhibiting what has been called "reporting bias."

One strategy for dealing with incomplete effect size data is to ignore the problem. This is clearly a bad strategy and is not recommended. If nothing else, such a strategy reduces the credibility of your meta-analysis because the presence of at least some missing data is obvious to most knowledgeable readers. Another problematic strategy for handling missing effect size data is to replace all of the missing values by some imputed value (usually zero). Although this strategy usually leads to a conservative estimate of the overall effect size, it creates serious problems in any attempt to study the variability of effect sizes and the relationship of study characteristics to effect size.

A better strategy in dealing with missing data is to extract from the study any available information about the outcome of the study. For example, you can often deduce the direction (sign) of the effect even when an effect size cannot be calculated. A tabulation of these directions of effects can then be used to supplement the effect size analysis (see e.g., Giaconia & Hedges, 1982; Crain & Mahard, 1983). You can even use such a

tabulation to derive a parametric estimate of effect (see Hedges & Olkin, 1980, 1985 for details on this procedure).

Perhaps the best strategy to deal with missing data on study outcomes is to use one of the many strategies that have been developed for handling missing data in sample surveys (see Madow, Nisselson, & Olkin, 1983). Generally, these strategies involve using the available information (including study characteristics) to estimate the structure of the study outcome data and the relationships among the study characteristics and study outcome. They can also be used to study the sensitivity of conclusions to the possible effects of missing data. Although these strategies have much to recommend them, they have been used only rarely in meta-analysis because they are difficult to implement. One example of the use of these methods in the context of validity generalization is given in Hedges (1987).

Example 12: Accounting for missing data

> Giaconia and Hedges (1982) reported the results of studies of the effects of open education programs (versus traditional education programs) on student self concept. They summarized the results of these studies by reporting the 84 effect size estimates that had a mean of 0.071 and a standard deviation of 0.418. However, they also reported the direction of the effect for a total of 100 independent comparisons in which 53 favored the open education group, 41 favored the traditional education group, and 6 comparisons could not be determined to favor either group.

2.2.2 Missing data on study characteristics

Another less obvious but equally critical form of missing data results from the incomplete descriptions of treatment, controls, or outcome measures. Missing data about study characteristics relate to the problem of breadth of study constructs and operations. If you attempt to code a high degree of detail about study characteristics, you will be faced with a greater degree of missing data when you code your studies. Yet, the alternative of coding vague study characteristics to ensure little or no missing data in your coding scheme is no less problematic. Neither procedure alone will inspire confidence among the readers of your meta-analysis.

One strategy for dealing with missing information about study characteristics is to have two levels of specificity: a broad level which can be coded for nearly all studies and a narrower level which can be coded for only a subset of studies. You will find this strategy useful if you exercise suitable care in describing the differences between the entire collection of studies and the smaller number of studies permitting the more specific analysis.

You can explore other alternatives to deal with missing data about study characteristics as well. One is a collection of relevant information from other sources such as technical reports, other descriptive reports on the program in the studies being examined, test reviews, or articles that describe a program, treatment, or measurement method under

review. The appropriate sources of this additional information are often published in the research reports being reviewed.

A second and often neglected source of information is the direct collection of new data. For example, in a meta-analysis of sex differences in helping behaviors, Eagly and Crowley (1986) surveyed a new sample of subjects to determine the degree of perceived danger in the helping situations examined in the studies. This rating was a valuable factor in explaining the variability of results across studies.

2.3 Publication bias

An excellent sampling plan cannot guarantee a representative sample if it is drawn from an incomplete enumeration of the population. The analogue in meta-analysis is that an apparently good sampling plan may be thwarted by applying the plan to an incomplete and unrepresentative subset of the studies that were actually conducted. This section discusses the problem of publication bias and ways to address that problem.

The published literature is particularly susceptible to the claim that it is unrepresentative of all studies that may have been conducted (the so-called publication bias problem). There is considerable empirical evidence that the published literature contains fewer statistically insignificant results than would be expected from the complete collection of all studies actually conducted (Bozarth & Roberts, 1972; Hedges, 1984). There is also direct evidence that journal editors and reviewers intentionally include statistical significance among their criteria for selecting manuscripts for publication (Greenwald, 1975; Bakan, 1966; Melton, 1962). The tendency of the published literature to over-represent statistically significant findings leads to biased overestimates of effect magnitudes from published literature, a phenomenon that was confirmed empirically by Smith's (1980) study of ten meta-analyses, each of which presented average effect size estimates for both published and unpublished sources.

Reporting bias is related to publication bias based on statistical significance. Reporting bias creates missing data when researchers fail to report the details of results of some statistical analyses, such as those that do not yield statistically significant results. The effect of reporting bias is identical to that of publication bias: Some effect magnitude estimates are simply unavailable.

One method for investigating the impact of publication or reporting bias is to compare effect size estimates derived from published (e.g., books, journal articles) and unpublished sources (e.g., conference presentations, contract reports, or doctoral dissertations). Be careful of such comparisons, however. The source of the study is usually confounded with many other study characteristics. An alternative method is to use statistical corrections for estimation of effect size under publication bias. This corresponds to modeling the sampling of studies as involving a censoring or truncation mechanism.

2.4 Establishing coding procedures

Once you have retrieved the studies for meta-analysis, you must establish procedures for translating critical study information into coded form amenable to computer processing. No comprehensive set of coding procedures could exhaust all of the study characteristics

that might be useful in any given meta-analysis. However, you might want to consider some study characteristics and coding procedures that have proven useful in other meta-analyses. These coding procedures and study characteristics are discussed in the next few sections.

2.4.1 Identifying contrasts within studies

You will need to devise a system of identification codes for referring to studies and to treatment-control (or other relevant) comparisons within studies. Computer based statistical analyses will treat each separate effect that may be analyzed as a "case." In some situations, there will be only one case (one effect size estimate) per study. But a single study may yield several cases (multiple effect size estimates) for two reasons: First, if there are several dependent variables of interest, an effect size estimate can be calculated for each, yielding one case for each dependent variable. Second, a study may yield several cases if it uses several independent samples of subjects such as students from several grade levels, SES categories, or school districts. When the primary study reports data for each of the independent samples, it is advisable to compute effect size estimates for each independent sample. If there are both multiple dependent variables and multiple samples, the total number of cases will be the product of the two. Thus, a study with two dependent variables and three samples will yield six effect size estimates. Computer files for data analysis are greatly facilitated by a coding scheme for identification numbers that completely distinguish each effect size estimate.

We suggest a coding scheme with identification numbers for each effect size estimate (case) that has the following components:

2.4.1.1 Study ID

This is a numeric code to identify the study from which the effect size estimate was obtained. This code should be cross-referenced to a listing of the bibliography of the meta-analysis.

2.4.1.2 Author reference

This is an alphabetic or character code which is usually the first few (8-16) characters of the primary study author's name and possibly the date of publication of the research report. This code serves as an easy way to identify the study in the bibliography.

2.4.1.3 Sample ID

This is a code to distinguish among the possibly many samples within a study. You need not construct the sample ID code to specify all of the details that may be needed to distinguish samples within a study (e.g., grade level, sex, SES, etc.). Separate codes should probably be used for each important characteristics of a sample.

Example 13: Coding subsample data within a study

> In a study that provided effect sizes for males and females separately at two grade levels, the four different samples might be given codes A, B, C, and D, but be distinguished by the other coded study characteristics of sex of sample and grade level of subjects.

2.4.1.4 Dependent variable

This is a code to reflect the dependent variable used to calculate the effect size estimate. Frequently, dependent variables will fall into related clusters. One broad construct such as mathematical achievement can be divided into subconstructs, such as mathematical computation, mathematics concepts, problem solving, etc. Moreover, each construct may be represented by several operations (specific test scales). Under these circumstances, you might find it useful to use this code to reflect the broad construct, the narrower construct, and the specific operation (test).

Example 14: Coding subtest or multiple measure data in a study

In a meta-analysis involving mathematics achievement, the codes 1000 to 1999 might indicate achievement, in general, while the codes 1100 to 1999 might reflect a particular test (e.g., 1123 might be the math problem solving scale of the Iowa Test of Basic Skills).

2.4.1.5 Time of testing

This is a code to indicate the temporal location of the measurement with the study.

Example 15: Coding temporal measurement characteristics in a study

Some studies have pretests, tests midway through the study, post-tests, and follow-up tests. Pretest data effect size estimates calculated from pretests scores should be coded whenever they are available since they are very useful in quantifying pre-existing differences between the treatment and the control group.

2.4.2 Study context

Study context factors include information describing the study, its setting, and its subjects. The details of context will vary with different meta-analyses. Following are some which might be considered in a meta-analysis of school- related research.

Example 16: Coding study context factors

- Location of the school (e.g., urban, suburban, rural)
- Type of school (e.g., parochial, public, private)
- Student population (e.g., SES, school size, gender)
- Teacher characteristics (e.g., selection, gender, educational background, experience, age, assignment)

2.4.3 Subject characteristics

The subject characteristics factor may be the most difficult of the study characteristics to code because most groups on which research is done are heterogeneous. But as in the case of educational studies, aggregate groups are often broken down into subgroups

which can easily be coded. Following is a list of subject characteristics which are common to most studies of students.

Example 17: Common subject characteristics

- Age or grade level
- Gender or predominant gender of groups
- Socioeconomic status
- Ability level
- Educational background
- Homogeneity of subjects (e.g., so the subjects represent any special groups such as talented and gifted, compensatory program subjects, etc.)

2.4.4 Study design and execution

One of the most important study characteristics is that of study design. Our experience is that study design features often account for more of the variation between study results than any other study characteristic. Therefore, you will want to make an extra effort to ferret out information about study design as you examine and code studies. Study design characteristics that we have found critical are discussed in the next four sections.

2.4.4.1 Controls for pre-existing difference

The first design characteristic you will want to examine and code is the type of control used to account for pre-existing differences between the treatment and control groups. The best design (for controlling pre-existing differences between groups) is to use random assignment of subjects to either treatment or control groups. However, this type of design is difficult to impose in educational research and other, less desirable designs are often used. One such design involves the random assignment of *intact classes*. Designs using random assignment of intact classes do not usually control for pre-existing differences as well as do designs involving random assignment of subjects. They should, therefore, be coded separately.

Sometimes subjects are randomly *sampled* from treatment or control classes. This may enhance representativeness of subjects but does not reduce bias due to pre-existing differences. Be careful not to confuse this procedure with the random assignment of subject design.

Explicit matching of subjects on a pretest that is highly correlated with the post-test is another design strategy that often produces relatively good control for pre-existing differences. Note however, that matching designs are weaker than random assignment designs because matching can only control for differences on the specific variables used in the matching. For example, matching on IQ controls for differences in ability but does not control for differences in motivation which might lead to biases in the results. Nonetheless, you will want to note studies using matching designs separately in your coding procedure.

Statistical control via the use of covariates or gain scores can also be used to minimize

the effects of pre-existing differences. But statistical control strategies suffer from the same weaknesses as matching designs; they control only for the variables specified. Finally, studies which use intact groups with no controls or with matching on only vague characteristics (such as SES) have the poorest control for pre-existing differences and should be noted in your coding scheme.

Regarding study design characteristics then, you will probably want to note the following types in your coding scheme:

Example 18: Design characteristics

- Random assignment of subjects
- Random assignment of intact classes
- Matching of subjects on pretest scores
- Matching of classes on mean pretest scores
- Cohort control (matching on SES, ability, etc.)
- No control
- No information

2.4.4.2 Experimental mortality

This is an indicator of how many subjects, classes, or schools dropped out of the study before post-testing. Even if the control for pre-existing difference between treatment and control subjects is initially excellent, attrition from either group introduces biases. Evidence of differential mortality is often hard to find in research reports, but if it is available, it can be helpful in explaining variations in the effect size aggregations performed later.

2.4.4.3 Treatment contamination of control groups

A problem arises when the control group actually receives more of the treatment than was expected at the outset of a study. Contamination arises, for example, when control group teachers learn of some aspects of the treatment and start using the strategies in their classes. Again, though this may be difficult to ascertain from a study report, it can prove very useful in explaining variations in subsequent effect size aggregations and is worth noting in your coding scheme.

2.4.4.4 Unit of analysis

This study characteristic refers to the source of the study statistics on which effect size estimates are computed. Most often statistics are based on subject scores. Occasionally however, study statistics are based on class, school, or some other group scores which tend to be less variable than the scores of individuals. Consequently, standard deviations of aggregated units will be systematically smaller than those of individual subjects. Because the magnitude of the standard deviation depends on the unit of analysis, effect size estimates (and correlation coefficients) based on different units of analysis are not comparable. That is, the very same subjects' scores yield different effect sizes depending on the unit of analysis used. Thus, you will want to note the unit of analysis used in your

coding procedure and aggregate studies using different units separately.

2.4.5 Treatment

A conceptually important source of variation between studies arises from variations among treatments. Because types of treatments may be qualitatively different across research domains, it is particularly difficult to describe how to characterize differences among treatment implementations. There are, however, some general guidelines that you may find helpful, but they are certainly not exhaustive.

2.4.5.1 Length of treatment

This factor simply describes how long the treatment was applied in the research study under review. It can vary widely between studies. For example, in the review of research on the effectiveness of new science curricula, treatment lengths from 2 to 36 weeks were observed.

2.4.5.2 Treatment fidelity

This factor is really many different factors combined under one descriptor—fidelity. It is an indicator of the degree to which the treatment is consistent with the theoretical descriptions of the treatment. One way to characterize treatment fidelity is to identify features or dimensions of the treatment that are theoretically important. In some cases, previous reviews or implementation studies may be helpful in arriving at dimensions useful in determining fidelity. In some cases, the research studies under review provide information from observation scales, interview protocols, or questionnaires which give evidence which is helpful in establishing level of treatment fidelity. Or you might devise some type of checklist to note the critical features of the treatment. Studies can then be grouped according to whether they have the particular constellation of features that correspond to the theoretical positions on the treatment.

This process of determining treatment fidelity sounds very abstract to this point. Perhaps an extended example from a study by Giaconia and Hedges (1982) will help to explain.

Example 19: Coding treatment fidelity

> In this study an attempt to characterize variability in implementations of open education was undertaken. The study began with a review of theoretical literature on open education. This conceptual analysis revealed four major features that were most central to theoretical conceptions of open education:
>
> 1. Role of the child in learning—referring to the extent that the child is active in guiding his or her own learning.
> 2. Diagnostic evaluation—the use of evaluation primarily for diagnostic purposes.
> 3. Materials to manipulate—the presence of materials to manipulate during the instruction.
> 4. Individualized instruction—presence of instruction based on the individual needs of each student.
>
> Because these features were considered most central to the theoretical conceptions

of open education, studies having all four characteristics were singled out for examination. The rationale for examining this group was that the implementations of open education in these studies had high fidelity to the theoretical conceptions of open education.

2.4.6 Control groups

Although control groups are often taken as self-explanatory, it is wise to remember that control groups can exhibit considerable variation. For example, an oft-used procedure in educational research is the "conventional instruction" control group. Conventional instruction is not a very specific construct, since most of the variation studied by educational research falls into this category. Variations within control groups are important because treatment effects are only defined with reference to their controls. A treatment that looks efficacious when compared with one type of control group may be far less efficacious when compared to another type of control group. Such differences in control groups produce substantial variability in treatment effect size.

Although research reports usually give little information about control groups, some major distinctions among control groups are useful to bear in mind.

Example 20: Specification of control group

- Conventional control—instruction or treatment which is already in place serves as the control
- No treatment—the control group receives no instruction or treatment related to the study (e.g., a study of a new science program where the control group receives no science instruction)
- Placebo—an alternative treatment that is not believed to be efficacious but is used to control for any "Hawthorne" effects
- Waiting list—control subjects are put on a waiting list and told that they will receive the treatment at a later date

2.4.7 Outcome variables

The conceptualization of outcome constructs and operations corresponding to those constructs is one of the most important activities in research synthesis. In addition to specifying outcome constructs precisely, there are other aspects of outcome variables that are usually useful to consider. These are outlined below:

- Congruence of outcome with treatment—the extent to which the outcome measure examines objectives that are likely to be affected by the treatment. For example, a test that measures a skill emphasized in the treatment but not mentioned in the control will bias the test scores and exaggerate the effect size. The reverse situation can also exist.
- Reactivity of the outcome measure—the extent to which the subjects' scores may depend upon what they believe the experimenter wants to hear. For example, self-reports of cooperative behavior may be more reactive than outside observations of group

behavior.

• Test and scale name—a code indicating the test, questionnaire, or observation scale and the subscale used in measuring a particular outcome variable.

• Format of outcome measure—an indication of the type of measurement; for example, oral, written, multiple-choice, free response, observational, etc.

2.5 Organizing data: Designing a coding sheet

Once you have decided the critical aspects of studies that should be coded, you are ready to design a coding sheet to facilitate the extraction and recording of this information from studies. The physical layout of a coding sheet is very similar to that of a questionnaire or interview protocol.

The actual coding sheet lists each item of information to be examined. Coding forms are usually designed to group pieces of information about a study that are logically related. It is usually convenient to categorize information into the groupings discussed earlier. Table 1 provides a checklist of coding information you may want to consider in designing your coding procedures.

Table 1: Coding Form Information Checklist

REFERENCE INFORMATION
- Complete reference (including form of publication)
- Study code (cross-listed with the bibliography)
- Coder identification number

SAMPLE CHARACTERISTICS
- Sample and subsample code
- Sample demographics (grade level, SES, ability level)
- School characteristics (size, type, location)

TREATMENT CHARACTERISTICS
- Type of treatment (as planned; as implemented)
- Theoretical dimensions of treatment
- Additional context variables (e.g., technology)

CONTROL GROUP CHARACTERISTICS
- Type of control group
- Theoretical dimensions of control group

TEACHER CHARACTERISTICS
- Years of experience, educational background
- Involvement with treatment (e.g., inservice training)

DESIGN CHARACTERISTICS
- Unit of analysis (individual or class)
- Control for pre-existing differences
- Experimental mortality
- Details of design

OUTCOME CHARACTERISTICS
- Type of criterion or outcome construct
- Congruence of treatments and outcome measure
- Method of measurement

EFFECT SIZE INFORMATION
- Source of effect size data for exact calculations
- Means
- Control group or pooled standard deviations
- F- or t-test statistics
- Repooled sums of squares within groups
- Samples sizes and sums of squares between groups
- Source of effect size data for approximate calculations
- ANCOVA adjusted means, sums of squares, and covariate post-test correlations
- Gainscore analyses and pretest-post-test correlations
- Direction of effect (sign of effect size)

It is usually useful to design the coding sheet so that the meaning of each item is clearly identified. You should make every effort to reduce ambiguity in specific categories. It is also advisable to over-specify categories at the coding stage rather than to under-specify. Categories can always be grouped at the analysis stage; they cannot be expanded without going back to the original studies. For example, an item on type of control for pre-existing differences between subjects in treatment versus control groups should specify the different procedures encountered in studies.

It is also useful to indicate a column number and a record (card) number for each data element so that the coded data can be entered directly from the coding forms. This avoids the problems of errors and expense associated with transcription of data to a new set of forms at data entry.

2.5.1 Sample coding sheet

The items on the coding sheet used in our study of the effects of new science curricula on student performance (Shymansky, Hedges, and Woodworth, 1987) are shown in Appendix A to illustrate some of the procedures discussed in earlier sections. Some problems associated with the original coding sheet have been corrected. You will note that this coding sheet specifies the column number to be used for each item so that it could be used directly by keypunchers for data entry. On the actual coding sheet these were written under the answer space. The first group of items (columns 4-27) refers to study context. Next is a series of study design and execution characteristics (columns 28-31, 38, 39). Then the treatment (curriculum) is coded (columns 43-46). This is followed by subject characteristics (columns 43-46), outcome variable characteristics (columns 47-48) and information used to compute the effect size estimate (columns 4-56).

2.5.2 Coding protocols and data screening

The use of coding sheets to extract information from studies should be done with great care to minimize errors in the coding process. An important and often neglected aspect of actually coding data is data screening. There are two aspects of data screening. One involves the search for values likely to be wrong as a result of transcription or data entry errors. The other involves the search for internal contradictions or unlikely values in the research reports themselves. Both aspects of data screening are vitally important to the quality of data produced by a meta-analysis.

When you are actually coding data from studies, it is useful to implement various checks for the consistency of the data reported in a study. Inconsistencies cast severe doubt upon the accuracy of data from a study since at least one of the two inconsistent values must be wrong. Inconsistencies frequently arise when the text of the research report and a table summarizing the results contradict one another. For example, the text of a study report might indicate that the treatment group outperformed the control group, but the corresponding table of means might indicate the opposite is true.

Another sort of internal inconsistency sometimes arises when studies report data in more than one way or in more than one table. You should check that the data agree across

tables. For example, the data listed as A, B, and C in Example 21 appeared in three separate tables of a reported study. Comparing the pretest means, standard deviations, and sample sizes reported in Part B of the table to those of Part C of the table, it is apparent that these data are contradictory—none of the post-test scores match.

Example 21: Contradictory data tables

		N	Mean	SD
A.	Pretest Scores			
	Experimental	27	62.63	10.00
	Control	24	66.29	9.42
B.	Post-test Scores			
	Experimental	27	60.37	7.80
	Control	24	62.75	8.75
C.	Pretest and Post-test Comparison			
	Experimental			
	Pretest	24	66.29	8.75
	Post-test	24	62.75	9.42
	Control			
	Pretest	27	62.62	7.80
	Post-test	27	60.37	10.11

Errors in transcription and data entry can also be identified by looking for inconsistencies among items in the coded data. It is also useful to search for unlikely or impossible values of each data element. For example, ages of subjects who are school age children usually lie between 5 and 19, grade levels should range between 0 and 12, sample sizes should be positive numbers of reasonable size, etc. A frequency distribution of each coded variable is often helpful in screening data for unlikely values. Joint distributions (crosstab tables) may also be useful in the search for particular combinations of variables that seem unlikely (e.g., a mother under 16 years of age with 8 or some other large number of children).

2.6 Reliability of coding

Once you have established coding procedures, you will need to demonstrate that your procedures can be applied reliably, i.e., that the procedures applied to the same study at different times by the same or different persons will produce the same coded values. "Intercoder" reliability estimates are usually determined by selecting a sample of studies which two or more coders independently code. A sample of about ten studies is normally used. The codes assigned by each coder are then compared to see how frequently they agree. It is advisable to examine the coding on an item-by-item basis. This will help you

identify items that are particularly difficult to code. Sometimes these items can be modified to make them easier to code to increase reliability. Sometimes the item has to be dropped because it cannot be coded reliably.

Although some researchers have set specific numbers for the minimum acceptable intercoder agreement, general guidelines are difficult. It is probably best to remember that reliability studies are designed as one (but not the only) test of data quality. Their purpose is to discover if the coded data are sound. Exactly how much intercoder reliability is necessary for a particular variable depends on the variable.

3.0 Data Analysis and Interpretation

Data analysis in quantitative research synthesis consists of defining a numerical index of study outcome and combining and studying the variation of these numerical estimates across studies. The usual numerical index of study outcome for studies of between-group differences is the standardized mean difference or effect size. Data analysis and interpretation begins with the computation of effect size estimates from the statistics reported in research studies. The variance of each effect size estimate is also computed. Preliminary analyses of the variation of effect sizes across studies are used to screen the data for outliers and possible coding errors. When coding errors have been corrected and obvious outliers have been deleted, the final analytic modeling of the effect size data can begin. Several different modeling strategies have been proposed for meta-analysis. The modeling strategy described herein is probably the most common. It uses so-called "fixed effects" models where possible and "random effects" models in cases where fixed effects fail to explain the variation of effect sizes among studies. In this chapter we will use examples from the meta-analysis of the effects of new science curricula to illustrate data analysis and interpretation.

3.1 Effect sizes and effect size estimates

How is the effect of treatment such as a new curriculum quantified? Imagine that traditional and new curricula could be taught to large groups of pupils drawn from the same population. At the end of instruction, each student would be assessed on one or more criterion measures. The effect of the new curriculum—how it compares with the old curriculum—will depend on the criterion measure used to assess it. For example, the new curriculum might have a substantial positive effect on pupils' attitude toward a science, but might not be any better than the old curriculum in imparting substantive knowledge of that science.

The *effect size* of the new curriculum is defined as the difference between the population mean criterion scores for new and traditional curricula expressed in standard deviation units. When the criterion measure is approximately normally distributed over the population, effect sizes are quite easy to interpret. For example, an effect size of 1.0 indicates that a pupil who would have been at the mean, or 50th percentile, under the old curriculum would be one standard deviation above the mean, that is at the 84th percentile, under the new curriculum. The new curriculum raises "C" students to "A" students. This example ought to suggest to an experienced teacher that effect sizes greater than 1.0 on achievement measures are to be viewed with skepticism, and may well reflect defects in the design and execution of a study.

3.1.1 Effect size estimates

In practice, effect sizes are estimated using samples from the relevant populations.The standard formula for estimating effect size is

$$g = \frac{\bar{x}_T - \bar{x}_C}{S} \qquad (1)$$

where $\bar{x}_T$ is the mean for the treatment group, $\bar{x}_C$ is the mean for the control group, and S is either the control group standard deviation or the pooled standard deviation.

It is intuitively reasonable to *estimate* the effect size using *sample* means and standard deviations in place of their population counterparts. However, there are *two* different sample standard deviations that could be used: one for the treatment group and one for the control group. There is also a pooled standard deviation that combines information from both groups. Which one should be used? The standard deviation of the control group has the advantage of being uncontaminated by any effects of the treatment. Consequently, it is often used to compute effect size estimates. Others prefer to use the pooled standard deviation which is slightly more stable as an estimate of the common standard deviation. To illustrate the procedures for calculating an effect size, data from a study by Aikenhead (1973) are used. The critical study elements and effect size calculation are shown in Example 22.

Example 22: Estimating effect size

Aikenhead described a study comparing the Harvard Project Physics (HPP) curriculum against traditional curricula. This study is unusual in that randomized assignment of teachers to treatment or control groups was employed. Aikenhead summarized the study this way:

"Fifty-five teachers were randomly selected from a total population of physics teachers in the United States and Canada. These teachers were then randomly assigned to teach *Harvard Project Physics* (after having participated in a summer institute) or non-HPP (the physics courses they would have ordinarily taught). An additional group of nineteen teachers, experienced at teaching HPP volunteered to participate in the evaluation project. They taught in various regions of the United States. A random sample of students of all teachers wrote the *Test on Understanding Science* (TOUS) or the *Science Process Inventory* (SPI) on a pretest and post-test basis . . . . There were 921 HPP and 267 non-HPP students."

The results are reported in the table on the next page:

Means And Standard Deviations

HPP	TOUS (N=445)	MEAN	STANDARD DEVIATION
	Pretest	34.43	6.857
	Post-test	37.54	7.059
	SPI (N=476)		
	Pretest	107.52	8.233
	Post-test	112.34	8.245
Non-HPP	TOUS (N=126)	MEAN	STANDARD DEVIATION
	Pretest	35.25	6.434
	Post-test	36.42	6.570
	SPI (N=141)		
	Pretest	107.08	7.789
	Post-test	109.30	9.481

Criterion Measure	Effect Size Estimate
TOUS	$\frac{(37.54 - 36.42)}{6.57} = 0.170$
SPI	$\frac{(112.34 - 109.30)}{9.48} = 0.321$

3.1.2 Correction for bias

The estimation procedure in Example 22 is essentially correct, although the resulting estimates are somewhat biased when the sample sizes are not large. It is, therefore, advisable to use the correction factor provided by Hedges and Olkin, Chapter 5, to produce an unbiased effect size estimator. The correction is a multiplier, J, which depends on the degrees of freedom for the standard deviation in the denominator of the effect size. For degrees of freedom above 50, the correction factor, J, is between 0.99 and 1 and can be ignored. Figure 2 gives formulas for raw (biased) and corrected (unbiased) effect size estimates.

FIGURE 2: Effect size computation

$\bar{x}_T$ is the mean and n_T the sample size for the control group

$\bar{x}_C$ is the mean and n_C the sample size for the treatment group

S is either the control group standard deviation or, if the meta-analyst prefers, the pooled standard deviation

m is the degrees of freedom of S, that is,

$m = n_C - 1$ if S is the control group standard deviation or

$m = n_C + n_T - 2$ if S is the pooled standard deviation.

The intuitive (biased) effect size estimate is

$$g = \frac{\bar{x}_T - \bar{x}_C}{S}$$

The correction factor for removing the bias of g is approximately:

$J = 1 - 3/(4m - 1)$

and the unbiased effect size estimate is the product of J and g,

$d = J \cdot g$

In the Aikenhead example the degrees of freedom are 125 for TOUS and 140 for SPI, since we used control group standard deviations to compute effect sizes. Consequently, the correction factors are 0.994 for TOUS and 0.995 for SPI yielding unbiased effect size estimates of 0.169 (0.170•0.994) for TOUS and 0.319 (0.321•0.995) for SPI. In this case the correction is hardly worth the bother, but it can be important in some cases. For example, if the control sample was only 20 students (19 degrees of freedom), the correction factor, J, would be (1 - 3/(4•19-1)) or 0.96, so that the unbiased effect size estimate would be 4% smaller than the uncorrected estimate.

3.2 Estimating effect size when means and standard deviation are not reported

Figures 2 and Example 22 of Section 3.1 show how to calculate effect size estimates when complete information is available about *both* means and standard deviations. Unfortunately, many research reports do not provide complete information and some ingenuity is usually required to calculate estimates of effect size from the information that is actually reported. The sections that follow present procedures to estimate effect size based on incomplete information.

3.2.1 Calculating effect size estimates when there are means and ANOVA tables, but no standard deviations

A common research design in education involves grouping subjects according to treatment and also by other categorizations such as age, sex, ability level, or grade level. The statistical analysis usually used for such designs is multifactor analysis of variance (ANOVA). Research reports frequently provide only the means for each cell and the ANOVA summary table. In this case it is still possible to calculate an estimate of effect size, but it is necessary first to compute the overall mean for treatment and control groups since the cell means will usually be means for subgroups of the treatment and control groups. It is also necessary to compute an overall standard deviation within the treatment and control groups.

Example 23: Estimating effect size from ANOVA data

Vanek (1974) described a small but elegant experimental study comparing the ESS curriculum with a textbook approach (Laidlaw Science Series),

> "Students from two existing third grades and two existing fourth grades were randomly assigned to the two . . . groups (i.e., ESS or traditional), at each grade level, so that approximately equal numbers of boys and girls were in each group. The two teachers at each grade level alternated, by units, teaching the ESS and the Laidlaw curricula . . . to eliminate teacher variables."

Vanek administered the Science Attitude Scale (SAS) to the pupils, along with other criterion measures. Her statistical report consisted of means and an analysis of covariance table for each criterion measure. Tables A and B are typical:

Table A (Adapted from Vanek's Table VII): Mean Scores of Science Attitude Scale, Total

		Boy	(n)	Girl	(n)
Grade 3	ESS	246.53	(15)	251.27	(11)
	Text	226.44	(16)	240.50	(12)
Grade 4	ESS	217.25	(16)	217.75	(12)
	Text	213.19	(16)	216.25	(12)

Table B (Adapted from Vanek's Table VI): Analysis of Variance Table, Science Attitude Scale, Total

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Grade	1	6204.20	6204.20	18.36
Curriculum	1	2486.24	2486.24	7.36
Sex	1	234.71	234.71	0.69
Grade•Curriculum	1	721.03	721.03	2.13
Grade•Sex	1	497.99	497.99	1.47
Curriculum•Sex	1	73.74	73.74	0.22
Grade•Curric•Sex	1	82.78	82.78	0.24
Residual	101	34125.95	337.88	
Total	108	44426.65		

Suppose that the cell means break the treatment and control group into m subgroups. The overall mean in the treatment groups is the weighted average of the individual treatment means,

$$\bar{x}^T = n_{1T}\bar{x}_{1T} + \ldots + n_m^T\bar{x}^T/(n_1^T + \ldots + n_m^T) \qquad (2)$$

where $n_1^T, \ldots, n_m^T$ are the sample sizes in each subgroup (cell) of the treatment group and $\bar{x}_1^T, \ldots, \bar{x}_m^T$ are the means of the subgroups of the treatment group. The overall mean in the control group is defined similarly,

$$\bar{x}^C = (n_1^C\bar{x}_1^C + \ldots + n_m^C\bar{x}_m^C)/(n_1^C + \ldots + n_m^C) \qquad (3)$$

where $n_1^C, \ldots, n_m^C$ are the sample sizes in each subgroup of the control group and $\bar{x}_1^C, \ldots, \bar{x}_m^C$ are the means in the subgroups of the control group.

Example 23: Estimating effect size from ANOVA data continued.

In Vanek's study, the treatment and control groups are broken into four subgroups because the treatment and control group are broken down by sex (male/female) and by grade (3/4). Thus m = 4 and the overall mean of the treatment group is

$$\bar{x}^T = \frac{15(245.53) + 11(251.27) + 16(217.25) + 12(217.75)}{15 + 11 + 16 + 12} = 232.15$$

$$\bar{x}^C = \frac{16(226.44) + 12(240.50) + 16(213.19) + 12(216.25)}{16 + 12 + 16 + 12} = 223.48$$

The overall standard deviation is obtained from the sums of squares and degrees of freedom presented in the ANOVA table. The pooled standard deviation S is

$$S = [SS_{balance}/(N - 2)]^{1/2} \qquad (4)$$

where $N = n_1^T + \ldots + n_m^T + n_1^C + \ldots + n_m^C$ is the total same size and

$SS_{balance} = SS_{total} - SS_{treatment}$. Thus, $SS_{balance}$ is the difference between the total sum of squares and the sum of squares for the treatment.

Example 23: Estimating effect size from ANOVA data continued.

In Vanek's data

$$S = [41940.41/(108 - 2)]^{1/2} = 19.89$$

The effect size estimate is then computed using $\bar{x}^T$, $\bar{x}^C$ and S as if the means and standard deviation were provided in the research report. That is

$$g = (\bar{x}^T - \bar{x}^C)/S,$$

the correction factor J is calculated as before and

$$d = J \cdot g.$$

Example 23: Estimating effect size from ANOVA data continued.

In Vanek's data

$$g = (234.42 - 223.48)/19.89 = 0.45$$

$$J = 1 - 3/(4\,106 - 1) = 0.993$$

$$d = J \cdot g = 0.45$$

Note that because the degrees of freedom are large the correction factor makes little difference.

3.2.2 Calculating effect size estimates when only t or F statistics are reported

Some studies fail even to report means, yet in many cases it is still possible to estimate effect sizes from test statistics which are reported. In some cases the effect size estimate can be calculated exactly as if means and standard deviations had been provided. In other situations, the effect size estimate can be only approximated.

Formulas for converting t or single-factor, one degree of freedom F into an effect size are given in section 3.2.2.1 Following that, we deal with multifactor analyses of variance or covariance and with adjusted means or t's in section 3.2.2.2.

3.2.2.1 Exact effect sizes from t or single-factor F

An exact effect size estimate can be derived either from a two-sample t statistic or an F statistic in a *single factor* analysis of variance in which the only factor is curriculum and exactly two curricula are compared (new vs. traditional). The unadjusted effect size, g, is

$$g = t \cdot \left(\frac{1}{n^T} + \frac{1}{n^C}\right)^{1/2} \qquad (5)$$

with degrees of freedom

$$m = n^T + n^C - 2$$

which would determine the bias adjustment J.

If F, rather than t is reported, then t can be recovered by the formula

$$t = (+/-)(F)^{1/2}$$

the sign of the square root is positive if pupils averaged higher under the new curriculum, otherwise it is negative.

3.2.2.2 Approximate effect sizes from multifactor ANOVA or ANCOVA

A substantial number of studies report analyses in which the effect of the new curriculum is adjusted for pre-existing differences between the treatment and control groups. The most common example is multifactor analysis of variance or covariance sometimes accompanied by covariate-adjusted means. Our approach in such situations is to approximate the effect size estimate as the difference between the adjusted means of new and traditional curricula divided by the pooled, *unadjusted* standard deviation, when available, otherwise by the adjusted standard deviation.

Suppose, for example, that Vanek had failed to report the means in Table 2 but had reported the multifactor ANOVA in Table B of Example 23. The key to unlocking the effect size in the absence of means is the Sum of Squares for curricula, $SS_{curric.}$ It turns out that this sum of squares is proportional to the squared difference between the adjusted means for pupils taught with the new curriculum and those taught with the old. These

means are, in effect, adjusted for all other factors (and covariates) in the analysis. (In Vanek's analysis, the means would be adjusted for grade and sex.) In other words, SS_{curric} supplies the numerator of the effect size estimate.

For the denominator of the effect size estimate we need the standard deviation of the criterion measure. One possibility is the square root of the residual mean square; however, we prefer to be somewhat conservative and include in the standard deviation all sources of variation in the criterion measure other than differences between curricula. Thus, we compute the standard deviation from the "balance" mean square.

Example 24: Estimating effect size from sums of squares data

In Vanek's ANOVA,

$$n^T = 55, \quad n^C = 56, SS_{curric} = 2486.24, \text{ and } SS_{balance} = 41940.41$$

with 107 degrees of freedom (there is some conflict between Vanek's reported n's and the total degrees of freedom in the ANOVA table). Consequently, the estimated effect size is

$$g = (+/-)((2486.24/41940.41)\bullet 107\bullet(1/55 + 1/56))^{1/2} = 0.48$$

which is very close to the effect size computed in section 3.2.1

The general procedure for computing effect sizes from studies reporting a multifactor ANOVA or ANCOVA can be summarized as consisting of several steps. First compute the difference between adjusted means as

$$\bar{x}^T_{adj} - \bar{x}^C_{adj} = \pm [SS_{treatment}(1/n^T + 1/n^C)]^{1/2} \qquad (6)$$

where n^T and n^C are the total number of pupils in the treatment and control group respectively. The appropriate sign (plus or minus) for the square root must be determined by reading the text of the report. If the author indicated that pupils did better under the new curriculum, then the sign is positive, if the opposite is true then the sign is negative. If there is no indication in the report one way or the other, then the effect size cannot be computed.

Next the pooled standard deviation is computed as

$$S = (SS_{balance}/m)^{1/2} \qquad (7)$$

where $m = n^T + n^C - 2$.

The effect size approximation (not bias-corrected) is

$$g = \frac{\bar{x}^{T}_{adj} - \bar{x}^{C}_{adj}}{S} = (+/-)(SS_{treatment}/SS_{balance}) \bullet m \bullet \left(\frac{1}{n^C} + \frac{1}{n^T}\right)^{1/2} \qquad (8)$$

which could be bias-corrected using factor J.

Example 25: Estimating effect size from multi-factor ANOVA or ANCOVA

Hipsher (1961) conducted a study that compared "the traditional high school physics curriculum and the curriculum developed by the Physical Science Study Committee (PSSC)."

"Four variables were statistically controlled: scholastic aptitude, prior achievement in natural science, physical science aptitude, and socio-economic status." The author reported that the adjusted mean for the PSSC group exceeded that of the control group by 9.5356, or "one-half the standard deviation of either group" (i.e., the effect size, g, is 0.50). The analysis of covariance summary table is shown below:

Analysis of Covariance Summary Table (adapted from Hipsher)

Source of Variation	Degrees of Freedom	Adjusted Sum of Squares
Curricula	1	4,269.4933
Residual	202	15,336.3417
Total	203	19,605.8350

Note that these sums of squares are all adjusted for the four covariates. The true total sum of squares, with 207 degrees of freedom (99+109-1), is not reported. Consequently, $SS_{balance}$ computed from this table will generally yield an underestimate of the standard deviation and, therefore, an overestimate of the magnitude of the treatment effect. Keeping this in mind, the effect size estimate is

$$g = ((4269.49/15336.34) \bullet 202 \bullet (1/99 + 1/109))1/2 = 1.04,$$

twice as big as the "one-half standard deviation" effect size reported by Hipsher.

This example makes it clear that the meta-analyst needs to use some caution in deriving

effect size estimates from covariate adjusted analyses. The largest bias resulting from the use of covariate adjusted sums of squares is in the underestimation of the standard deviation of the criterion measure. The appropriate correction factor for this bias turns out to be $(1 - R^2)^{1/2}$, where R is the multiple correlation between the criterion measure and the covariates. The corrected raw effect size is

$$g_{corrected} = g \bullet (1-R^2)^{1/2}. \qquad (9)$$

In Hipsher's case the correction factor appears to be about 0.5 corresponding to an R^2 of 0.75, since he reported an effect size of 0.5 and we computed an effect size of 1.04. We were able to deduce the correction factor in this study because Hipsher reported the effect size. This is quite unusual; in general the effect size is not reported and the meta-analyst must use collateral information to estimate the multiple correlation between the criterion measure and the covariates.

Our practice in the meta-analysis of the effects of new science curricula, which is certainly open to debate, has been to assume an R^2 of 0.5; i.e., to multiply g by 0.7 when it is derived from covariate adjusted sums of squares. Of course, when the unadjusted total sum of squares is reported, there is no need to correct the effect size estimate since the pooled standard deviation can be estimated from the "balance" sum of squares.

If adjusted means are reported as well as an ANCOVA table, the procedure illustrated in Example 25 should be used, followed by a $(1 - R^2)^{1/2}$ correction.

The formulas and examples we have presented so far are incorporated in the SAS program in Appendix III. Although they do cover the majority of useful cases, the meta-analyst will find many unique combinations of statistical information not covered here. Our best advice in those cases is to consult a statistician.

3.2.3 Sometimes effect sizes cannot be estimated

Finally, here are some examples in which effect sizes cannot be estimated from reported data.

3.2.3.1 Paired designs

Some studies match pupils in the control and treatment groups. This is appropriate for reducing bias in statistical inference; however, if separate standard deviations are not reported, the effect size cannot be computed. One study, for example, presented the matched pairs t-test shown in Example 26.

Example 26: A "paired" design in which effect size cannot be calculated

Group	N	Mean	Sum of Differences	Sum of Squared Differences	t
Control	63	14.27			
			-87	1099	-3.90
Experimental	63	12.32			

Unfortunately, the author didn't report individual standard deviations, so the method of Figure 2 can't be used. The method shown in Example 25 is not valid here since it requires an "unpaired," independent samples t-test. In short, the effect size cannot be estimated without additional information (see McGaw and Glass, 1980).

3.2.3.2 Insignificant differences

Studies which report their findings in the form of analysis of variance tables but fail to report means and standard deviations frequently also fail to state whether insignificant differences are positive or negative. While it could be argued that this is standard practice not to interpret individual insignificant results, it should be kept in mind that a series of individually insignificant differences can add up across studies to a significant difference *if they are all in the same direction.* To make an analogy, heads on a single coin toss wouldn't make an impression, but twenty heads in a row would.

3.3 Standard error of effect size estimates

The sampling standard error of an effect size estimate is the standard deviation of the estimated effect size around the true effect size in the population of students from which the study population was selected. Sampling standard error measures the sampling variation of the estimated effect size but does not reflect non-sampling variations which would occur if the study had used a different population of students or different teachers.

A very accurate approximation to the sampling variance (the square of the sampling standard error) of the effect size estimate d is

$$S_d^2 = \frac{n^T + n^C}{n^T n^C} + \frac{d^2}{2m} \quad . \qquad (10)$$

The exact formula for sampling variance of d is

$$S_d^2 = J^2 \quad \frac{m}{(m-2)} \left(\frac{n^T + n^C}{n^T n^C} \right) + d^2 \left(\frac{J^2 m}{m-2} - 1 \right) . \qquad (11)$$

But this exact formula is seldom needed. The approximate formula is accurate enough for all meta-analyses except those in which the degrees of freedom are small (i.e., m <10).

Example 27: Estimating sampling standard error

In the Aikenhead study on the TOUS scale the sample sizes were $n^T = 445$ and $n^C = 126$; degrees of freedom, m, were 125; the bias correction, J, was 0.994; and the corrected effect size, d, was 0.169 (see Figure 2 and the paragraph following it). Consequently, the sampling variance of d given by the appropriate formula is

$$S_d^2 = \frac{445 + 126}{445 \cdot 126} + \frac{(0.169)^2}{2(445 + 126)} = .0102.$$

Computation of the sampling variance using the exact formula gives

$$S_d^2 = 0.994^2\left(\frac{125}{123}\right)\left(\frac{445 + 126}{445 \cdot 126}\right) + 0.169^2 \cdot 0.994^2 \left(\frac{125}{123} - 1\right) = 0.0102.$$

Thus, in this case the two formulas agree to four decimal places.

Why are we interested in the sampling standard error of the effect size? The full answer will become clear later, but perhaps a hypothetical example will shed some light.

Suppose that there are two independent studies comparing ESS with a textbook oriented curriculum for third grade boys. Say that one study produced an effect size of 0.37 with a standard error of 0.50 while the other, larger study yielded an effect size of 0.31 with a standard error of 0.16. Although neither effect size by itself is statistically significant (each is less than two standard errors from zero), when they are properly combined the effect is statistically significant.

It turns out that the statistically optimal way in which to combine two effect size estimates is to compute their weighted average, weighted in proportion to the reciprocals of the squares of their variances, that is,

$$d_{combined} = \frac{0.37(1/0.50^2) + 0.31(1/0.16^2)}{(1/0.50^2) + (1/0.16^2)} = 0.32.$$

In this case, the combined effect size is more than two standard errors away from zero, since the standard error of this combined effect size estimate is the square root of the reciprocal of the sum of reciprocals of the individual variances

$$S_{d(combined)} = \frac{1}{((1/.50)^2 + (1/.16)^2)^{1/2}} = 0.152$$

In general, if most studies yield effect size estimates favoring the new curriculum, their combined effect size will be more statistically significant than any individual study.

In short, the sampling standard errors of individual effect size estimates provide weights for optimally combining effect sizes across studies and in addition provide information for computing standard errors of combined effect sizes.

Another issue which can be addressed by using sampling standard errors is the

question of effect heterogeneity, which we will discuss in detail later. In our hypothetical example the heterogeneity issue boils down to the question of whether two different studies of the same curriculum produced significantly different effect sizes. In this example, the answer is no, since the effect sizes of 0.37 and 0.31 are within one standard error of each other. When there are significant differences in the effect sizes found in different studies, the meta-analyst would search for study characteristics to explain the difference using an ANOVA analog called Analysis of Heterogeneity.

3.4 Combining effect size estimates

One goal of meta-analysis is to combine estimates of effect size to produce an overall average. There are, however, two somewhat different situations in which effect size estimates are combined. One situation is that of combining independent effect size estimates across studies. The other situation arises when correlated effect sizes are combined within studies. Each situation is discussed separately below.

3.4.1 Combining independent effect size estimates by weighted averaging

Several effect size estimates, obtained from studies of a particular curriculum, conducted under similar conditions with similar pupil populations and similar criterion measures, can be combined to give an overall effect size estimate. As we said earlier, the power and sensitivity of meta-analysis comes from the fact that this combined estimate will have smaller standard error than any of its parts. The statistically optimal way to average a group of independent estimates is to form a weighted average, weighing each estimate by the reciprocal of its sampling variance, i.e., the reciprocal of its squared standard error. Figure 3 exhibits the formula for this calculation.

FIGURE 3

Suppose that

$$d_1, d_2, \ldots, d_k$$

are k independent effect size estimates and that their standard errors are,

$$S_1, S_2, \ldots, S_k.$$

Then the weighted average effect size is

$$d = \frac{d_1/S_1^2 + d_2/S_2^2 + \ldots + d_k/S_k^2}{1/S_1^2 + 1/S_2^2 + \ldots + 1/S_k^2}. \qquad (12)$$

The standard error of the weighted average is

$$S_{d_+} = 1/(1/S_1^2 + 1/S_2^2 + \ldots + 1/S_k^2)^{1/2}. \qquad (13)$$

Example 28: Combining weighted effect size estimates

If three studies yield effect sizes of 0.10, 0.35, and 0.60 with standard errors of 0.2, 0.1 and 0.3, respectively, then the weighted average effect size is

$$d_+ = \frac{(0.10/0.04) + (0.35/0.01) + (0.60/0.09)}{(1/0.04) + (1/0.01) + (1/0.09)} = 0.324,$$

and its standard error is

$$S_{d+} = 1/[(1/0.04) + (1/0.01) + (1/0.09)]^{1/2} = 0.086.$$

The most precise estimate (0.35) receives the greatest weight. The standard error of the weighted average is smaller than that of any of its components.

3.4.2 Combining correlated effect size estimates

The goal of meta-analysis is to combine information from several studies. One difficulty, however, is the great variety of criterion measures used by different investigators: for example, one investigator might measure science achievement by a standardized test (e.g, *Sequential Test of Educational Progress* (STEP)), another might write his/her own test, a third might use yet another test (e.g., *Test on Understanding Science* (TOUS)). In order to have adequate numbers of studies for meta-analysis it is necessary to combine and compare effect sizes for different criterion measures. Of course, it would not be appropriate to compare or combine criteria measuring different concepts like attitude changes and achievement. It would not be meaningful, for example, to ask if a new curriculum changed male attitudes toward science more than it changed female understanding of science. On the other hand, it is reasonable to combine or compare effect sizes for similar criterion measures such as two different achievement tests. For these reasons, we grouped criterion measures into five criterion clusters (Achievement, Perceptions, Process Skills, Analytic Skills and Other Performance Areas) in our study of new science curricula. This, however, introduces a technical statistical problem. In many cases, two or more different criterion measures within one study fall into the same criterion cluster. For example, Wideen (1971) reported means and standard deviations for six criterion measures (Example 29). The first three (II, RAI and TLE) measure *perceptions*, the next two (PPMA and PPMB) measure process skills, and the last (STEP-Science) measures achievement. Since these measures were all made on the same groups of pupils, they are statistically correlated. It would be wrong to regard them as three independent estimates of the effect of the new curriculum.

How then do we deal with correlated effect sizes within the same study (e.g., the three perceptions effects, or the two process skills effects in Example 29)? (Hedges and Olkin 1985,[Chapter 10, Section G]) recommend selecting one representative effect size from each cluster, perhaps at random. Another alternative is to average the correlated effect

size estimates where they occur and make a conservative estimate of the standard error of their average as shown in Table B of Example 29. For example, the average effect size for the "perceptions" cluster is (-0.174+0.348+0.016)/3 or 0.063 and a "conservative" standard error of this average is the average of the individual standard errors: (0.085+0.086+0.085)/3 or 0.085. "Conservative" in this context means that the true standard error (which we lack information to estimate) is somewhat smaller than 0.085. (See Appendix 2 for an explanation.)

Example 29: Combining correlated effect size estimates

Table A: Means and Standard Deviations for Experimental and Control Student Groups for Criterion Measures (adapted from Wideen, 1971, Table VII)

Criterion Measure	Experimental (n=263) Mean	SD	Control (n=292) Mean	SD	Effect Size Esti. mate	Std. Err.
Perceptions						
Interest Inventory (II)	26.97	9.91	28.52	8.90	-0.174	0.085
Revised Attitude Inventory (RAI)	67.24	7.86	64.39	8.18	0.348	0.086
Student Survey of the Teaching-Learning Environment (TLE)	66.15	14.84	65.93	13.33	0.016	0.085
Process Skills						
SAPA Pupil Process Measure A (PPMA)	25.05	7.84	17.46	7.90	0.951	0.094
SAPA Pupil Process Measure B (PPMB)	12.80	3.74	10.69	4.31	0.490	0.087
Achievement						
Sequential Test of Educational Progress: Science (STEP-Science)	67.56	25.43	60.43	25.53	0.279	0.086

After correlated effect size estimates are averaged together to form Table B, the data contain at most one effect size estimate for each criterion cluster for each independent

subgroup of subjects. Of course, there will still be studies with two or more correlated effect sizes but these will be in different criterion clusters—will measure different constructs. Since we analyze each criterion clusters separately and never make comparisons between criterion clusters, the comparisons we do make (for example, boys' achievement vs. girls' achievement) are between independent groups of subjects.

Table B: Aggregate Effect Sizes for Criterion Clusters within Wideen's (1971) Study

Criterion Cluster	Average Effect Size	Conservative Standard Error
Perceptions (II, RAI, TLE)	0.063	0.085
Process Skills (PPMA, PPMB)	0.723	0.091

It is still possible for one study to produce two or more independent effect size estimates in the same criterion cluster. This occurs when the investigator reports statistics for independent subgroups of subjects. For example, Vanek (1974), reported statistics for third and fourth grade girls and boys, which yielded four independent effect size estimates. (See Example 23.)

To summarize, the data for each individual study are reduced as follows: for each independent subgroup of pupils, effect sizes within the same criterion cluster are averaged together. Appendix 3 contains SAS programs for carrying out these calculations. Effect sizes computed by these programs for achievement criterion measures are listed in Appendix 4.

3.5 When is it appropriate to combine estimates: Measuring heterogeneity

Independent studies may produce effect size estimates which differ by many times their sampling standard errors. Appendix 4, for example, displays effect size estimates based on achievement tests for various experimental science curricula. The range of effect sizes even within the same curriculum is striking. For example, effect sizes for BSCS Yellow vary from about -0.8 to +0.8, as shown in Example 30.

Example 30: Examining heterogeneity

BSCS Yellow—Effect Size Estimates and Standard Errors

Study ID	Effect Size (d)	Standard Error (S)	Notes on the Study
9A	0.67660	0.069420	Israel-city
9B	0.93613	0.92565	Israel-kibbutz
9C	0.03495	0.090629	Israel-ag school
29	0.44184	0.96960	boys
37A	0.28069	0.140289	zoology classes
37B	0.52985	0.142355	botany classes
37C	0.38884	0.141362	biology classes
43	0.56643	0.165552	
51	0.21218	0.061099	
63	0.71001	0.149950	
65	0.63274	0.186029	passive control
82A	0.79051	0.475224	BSCS inquiry, boys
82B	0.44143	0.340299	BSCS inquiry, girls
82C	0.02289	0.392769	BSCS traditional, boys
82D	-0.81212	0.353206	BSCS traditional, girls

The variation among studies is, of course, due in part to random sampling fluctuations as reflected in the sampling standard errors. However, in some cases differences between individual studies exceed several standard errors, presumably reflecting differences in the characteristics of those studies. In Example 30, the highest and lowest effect sizes came from studies 9 and 82, both of which have unusual characteristics. Study 9 was conducted in a non-English-speaking culture and the deviant effects in study 82 occurred when traditional teaching methods were used with the experimental curriculum. However, even with these unusual studies set aside there remain substantial differences among the remaining studies. Therefore, it appears that effect sizes are influenced by study characteristics not captured by the variables recorded by the meta-analyst. To study this 'non-sampling' variation we use heterogeneity analysis.

The fundamental measure of heterogeneity, Q, is based on the idea that the expected squared deviation of an effect size estimate from its true value equals the square of its standard error (Hedges and Olkin 1985,[Chapter 6, Section D]). The formula for the Q statistics is given below.

Suppose that

$$d_1, d_2, \ldots, d_k$$

are k independent effect size estimates with standard errors,

$$S_1, S_2, \ldots, S_k,$$

and weighted average effect size d_+.

The heterogeneity statistic is

$$Q = ((d_1-d_+)/s_1)^2 + ((d_2-d_+)/s_2)^2 + \ldots + ((d_k-d_+)/s_k)^2 \qquad (14)$$

$$= Q_1 + Q_2 + \ldots + Q_k$$

Example 31: Calculating the heterogeneity (Q) statistic

Using the data in Example 30, the weighted average effect size is 0.4415, and the heterogeneity statistic is

$$Q = ((0.6766-0.4415)/0.0694)^2 + \ldots + ((-0.8121-0.4415)/0.3532)^2$$

$$Q = 11.47 + 28.56 + 20.12 + 0.00 + 1.31 + 0.39 + 0.14 + 0.57 + 14.09 + 3.21 + 1.06 + 0.54 + 00 + 1.14 + 12.60$$

$$Q = 95.18$$

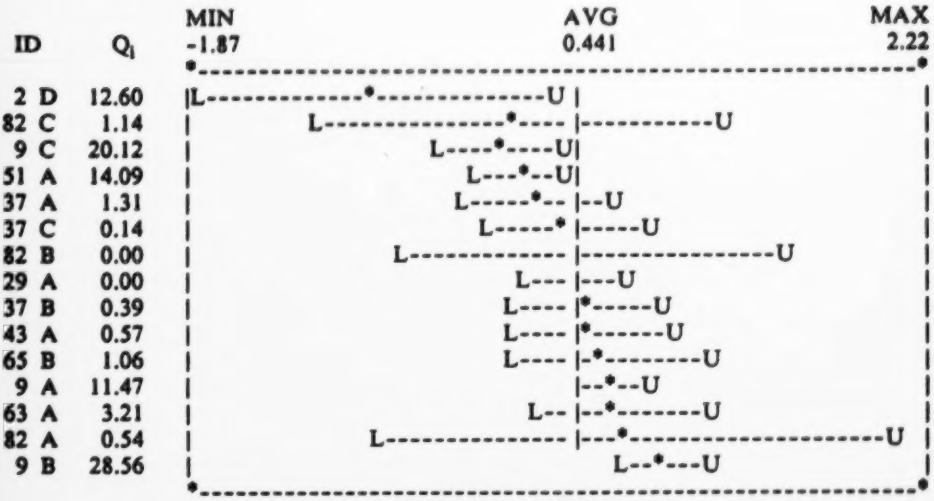
A test for heterogeneity (i.e., variation in excess of sampling fluctuations) is carried out by comparing Q to a percentile of the chi-squared distribution with k - 1 degrees of freedom. For the data in this example, the Q statistic (95.18) exceeds the 99th percentile of the chi-squared distribution with 14 degrees of freedom (29.14); consequently, the effect sizes for BSCS Yellow shown in Example 30 are significantly heterogeneous.

The individual squared deviations (11.47, 28.56, . . . , 12.60) are denoted $Q_1, Q_2, \ldots, Q_k$. These Q_i values are useful for identifying studies which deviated significantly from the weighted average. As a rule of thumb, only one study in 19 should have a Q_i value exceeding 4 and only one in one hundred should exceed 5.3. Thus, effect sizes 9A, 9B, 9C, 51, and 82D, with Q_i values greatly exceeding 5.3, are strikingly deviant from the other ten, and the meta-analyst would be well advised to examine the relevant studies to determine, if possible, why they are deviant.

3.5.1 Heterogeneity plots

A convenient way to screen a group of studies for heterogeneity is to plot their effect sizes with error bars of plus or minus two or three standard errors as shown in Figure 4 for studies of the BSCS Yellow curriculum in Example 30. Notice that the error bars of deviant effects fail to cover the weighted average (0.441). This figure was produced by SAS program III listed in Appendix 3. Heterogeneity plots and Q_i statistics for the other new science curricula are presented in Appendix 4.

FIGURE 4: Heterogeneity plot for BSCS Yellow achievement effects



Key: U = (Effect size + 3 sampling standard errors)
* = (Effect size)
L = (Effect size - 3 sampling standard errors)
| = (Weighted average effect size for studies of this curriculum)

ID = Study Identification code.

$Q_i = ((d_i - d_+)/s_i)^2$ (Squared deviation of ith effect size from the weighted average)

3.6 Formal analysis of heterogeneity: An analysis of variance for effect sizes

Given that a group of effect sizes is significantly heterogeneous (has a significant Q statistic), it may be possible to explain the excess variability in terms of observed characteristics of the studies which yielded the effect size estimates. The question is investigated using weighted analysis of variance with effect size as the dependent variable. The reciprocal, squared sampling standard errors are used as weights, and the factors in the analysis of variance are those study characteristics which the meta-analyst suspects may account for variation among studies.

The table given in Example 32 is illustrative of this type of weighted ANOVA. This particular example analyzed all achievement effect sizes using Curriculum, IQ and sex as factors. The analysis included only main effects and the IQ x Sex interaction, although other interactions could have been included. The ANOVA table shows that even though much of the total heterogeneity can be explained by Curriculum, IQ and Sex, there still remains significant unexplained heterogeneity.

The choice of IQ and sex as explanatory factors is somewhat arbitrary, since these variables are highly collinear with other potential explanatory variables. Indeed, it is rarely possible to include more than one or two factors in a formal analysis of heterogeneity due to widespread confounding (collinearity) of potential explanatory variables—meta-analyses are not experiments and are not 'balanced' with respect to factors of interest. Formulas for weighed ANOVA are to be found in (Hedges and Olkin 1985,(Chapter 7). SAS program III in Appendix 3 produced the table in Example 32, and it can be adapted to carry out most weighted ANOVAS of interest.

Example 32: Analysis of heterogeneity table for achievement effect size

Source of Heterogeneity	Q	Degrees of Freedom	Significance
Curricula	324.6	17	.01
IQ	16.9	2	.01
Sex	0.3	1	ns
IQ x Sex	36.3	1	.01
Unexplained	466.8	55	.01
Total	821.0	80	.01

3.7 Combining effect size estimates with unexplained heterogeneity: Random effects models in meta-analysis

Random effects procedures for combining estimates in meta-analysis differ from the fixed effects procedures described in sections 3.4, 3.5, and 3.6 in that they treat between-study variations in effect sizes as random. In the random effects model, the studies that

are actually performed are viewed as a sample from a universe of possible studies that might have been performed. The population effects sizes (the effect size parameters) for the studies are treated as a sample from a universe (or hyper-population) of possible effect sizes. Thus, there are two sources of between-study variation in the observed sample effect size estimates. One source of variation is the between-study variation in the underlying effect size parameters. The second source of variation is the sampling error of the observed effect size estimate about its underlying effect size parameter. Random effects analyses take into account the between-study variation in effect size parameters by formally estimating the magnitude of this variation via a *variance component*. (A complete discussion of the rationale and methods for random effects analyses is given in Chapter 9 of Hedges and Olin (1985).)

3.7.1 Estimating the variance component

The between-study variance component is essentially the amount by which the observed variance among effect size estimates exceeds the within-study sampling variance. Consequently, the variance component σ_δ^2 is usually estimated as the difference between the observed variance among effect size estimates and the average of the sampling error variances. If $S^2(d)$ is the usual sample variance of $d_1,...,d_k$ and are the sampling variances of $d_1,...,d_k$ given in section 3.3, the variance component estimate $\hat{\sigma}_\delta^2$ is

$$\hat{\sigma}_\delta^2 = S^2(d) - (S_1^2 + ... + S_k^2)/k.$$

Example 33: Estimating sampling variance

If three studies yield effects of 0.10, 0.35, and 0.60 with standard errors of 0.2, 0.1, and 0.3, respectively, then the unweighted average of the effects in $\bar{d} = 0.35$ usual sample variance of the effects is

$$S^2(d) = \frac{(0.10-0.35)^2 + (0.35-0.35)^2 + (0.60-0.35)^2}{2} = 0.0625,$$

and the variance component estimate is

$$\hat{\sigma}_\delta^2 = 0.0625 - (0.04 + 0.01 + 0.09)/3 = 0.0158.$$

This suggests that the distribution of the random effects has a variance of 0.0158 or a standard deviation of about 0.13.

Note that a test of the hypothesis that $\sigma_\delta^2 = 0$ is equivalent to the heterogeneity test for fixed effects models given in section 3.5. This is because $\sigma_\delta^2 = 0$ if and only if all of the studies have the same population effect size. Consequently, a test that all of the population effect sizes are the same is also a test that $\sigma_\delta^2 = 0$. It is important to recognize that the

estimate $\hat{\sigma}^2_\delta$ of σ^2_δ can differ substantially from zero even if it is not large enough to be statistically significant. Moreover, a statistically significant value of σ^2_δ need not be large in an absolute sense. Consequently careful judgment is required in the interpretation of the variance component.

3.7.2 Combining effect size estimates in random effects models

Effect size estimates are combined in random effects models by computing a weighted average in a manner similar to that in fixed effects models. The only difference is in the definition of the weights. In random effects models both within-study sampling error variance (the standard error) and the variance component contribute to the weights.

Suppose that $d_1,....,d_k$ are k independent effect size estimates with standard errors

$$S_1,....,S_k,$$

and that the variance component estimate is $\hat{\sigma}^2_\delta$. Then the random effects weighted average is

$$d = \frac{d_1/(S_1^2 + \hat{\sigma}^2_\delta) + ... + d_k/(S_k^2 + \hat{\sigma}^2_\delta)}{1/(S_1^2 + \hat{\sigma}^2_\delta) + ... + 1/(S_k^2 + \hat{\sigma}^2_\delta)} \qquad (15)$$

The standard error of the random effects weighted average is

$$S_d = 1/[1/(S_1^2 + \hat{\sigma}^2_\delta) + ... + 1/(S_k^2 + \hat{\sigma}^2_\delta)]^{-1/2}. \qquad (16)$$

Example 34: Combining effect size estimates in a random effects model

If three studies yield effect sizes of 0.10, 0.35, and 0.60 with standard errors of 0.2, 0.1, and 0.3, respectively, the variance component estimate is $\hat{\sigma}^2_\delta$ = 0.0158. The weighted average effect size is

$$d = \frac{(0.10/0.0558) + (0.35/0.0258) + (0.60/0.1058)}{(1/0.0558) + (1/0.0258) + (1/0.1058)} = 0.318$$

and the standard error is

$$S_d = 1/[(1/0.0558) + (1/0.0258) + (1/0.1058)]^{1/2} = 0.123.$$

Note that the weighted mean effect under the random effects model d = 0.318 is slightly smaller than the weighted mean effect under the fixed effects model $d_+ = 0.324$. Note also that the standard error $S_{d+} = 0.086$ under the fixed effects model is smaller than the standard error $S_d = 0.123$ under the random effects model.

Note that the weights used in the random effects model are not the same as those of

the fixed effects model discussed in section 3.4 unless the variance component estimate $\hat{\sigma}^2_\delta$ is exactly zero. Because $\hat{\sigma}^2_\delta$ is usually larger than zero, the weights are generally smaller in the random effects case and d usually differs from d_+. Moreover, the standard error in the random effects case is usually larger (often much larger) than in the fixed effects case. As a result, overall average effect sizes that are significantly different from zero (i.e., more than two standard errors away from zero) in a fixed effects analysis may not be significant in a random effects analysis. The difference, of course, results from differences in the conceptualization of the model, and in what counts as random. The SAS program in Appendix III computes both fixed and random effects estimates of effect sizes and their standard errors.

4.0 Reporting Results

Effective reporting of the results of a meta-analysis requires considerable care and good judgment. The report must be complete enough to describe clearly what was done but concise enough to be readable. It must provide a context in which to interpret results and link them to other theory and empirical results. One useful overall guideline is that the report of a meta-analysis should be organized like that of any empirical research report. It should begin with a clear statement of the problem, describing the constructs and their operations. If the meta-analysis is oriented toward hypothesis testing, the hypothesis should be explicitly stated. In meta-analyses devoted to hypothesis generation, the range of hypotheses to be explored should be specified as clearly as possible. Procedures for data collection such as the procedures for identifying and sampling of studies should be described in detail. Similarly, the procedures used for extracting information from studies (coding of study characteristics) and procedures used to insure the quality of these data (such as reliability checks) should be described. In particular, missing data and the reasons that they are missing should be identified. Procedures for data evaluation, such as ratings of study quality should be described in detail. It is probably a good idea to describe the criteria used in ratings of study quality and to provide justification for them.

It is also helpful to describe a few key studies in detail to help readers develop a clearer intuitive understanding of the research. The clinical discussion of particularly important studies or of studies that yield discrepant findings is also useful. For example, if only one or two studies examined the interaction of treatment with a potentially important variable such as ability level, it might be wise to discuss those studies and the implications of their findings for general conclusions about treatment effects.
Finally, a long table providing a summary of the critical aspects of each study is also useful to present a broad picture of the data available. Such a table might present, for each study, a brief description of the independent variable (e.g., treatment and control), the dependent variable, crucial features of the study design (such as sample size, type of assignment of subjects to treatments, etc.), the result as reported by the original investigator, and the estimate of effect size. When the number of studies is large such a table may be many pages long and some journals may be reluctant to publish it. However, it will greatly increase the credibility and usefulness of your analysis and make it available to individuals who request it even if it cannot be published in its entirety. An example of such a summary table is shown in Example 35 (page 49). The table is an excerpt from a long table summarizing study characteristics and outcomes. It is adapted from Table 3 of Eagly and Crowley (1986) which provided data on 99 studies.

The presentation of the data analysis should include enough information to make analyses interpretable. The type of analysis (fixed or random effects) should be described and the relevant summary statistics should be provided. When a mean effect size is presented, its standard error should always be given. It is often useful to provide both fixed effects standard errors (that is, S_d) and random effects standard errors (that is, S_d) for means. In addition, the homogeneity statistic, a variance component estimate, or both

should be given as an index of variability of effects across studies. When either the homogeneity statistic or the variance component suggest heterogeneity among study results, heterogeneity plots like those in Figure 4 can be useful in interpreting results. Finally, when categorical models are used to explain variability among effect sizes via study characteristics, an overall summary table should be presented along with a table of cell means with standard errors and some indication of variability of effects within cells (like a homogeneity statistic or variance component estimate).

The final indispensable element of a good report of a meta-analysis is a thoughtful reflective discussion of the findings. The discussion should link the results to the broader context of other findings, intuitions, and common sense. It should show how the findings of this meta-analysis make sense and should show implications for future research and for practice.

Example 35: A Meta-Analysis Summary Table

Sex-of-Subject Effect Sizes and Study Variables, Ordered by Magnitude of Effect Sizes

Study	Behavior[a]	Effect size (d)[b]	95% CI for d (lower/upper)	Categorical variables[c]	Sex differences in judgments of helping behaviors: Competence[d]	Comfort	Danger	Own Behavior	Stereotypic
1. Pomazal & Clore (1973), Study 3	Helping a person with a flat tire or picking up a hitchiker	1.48 (0.34/0.03)	1.29/1.66	3/3/2/2	0.96	0.59	0.69	0.72	1.37
2. I.M. Piliavin, J.A. Piliavin, & Rodin (1975)	Helping a man who fell in the subway	1.44 (0.10/0.00)	1.32/1.56	3/3/2/2	0.52	-0.03	0.34	0.23	0.00
3. Pomazal & Clore (1973), Study 1	Helping a person with a flat tire	1.44 (0.21/0.01)	1.22/1.66	3/3/2/2	1.11	0.26	0.72	0.56	1.35
4. Pomazal & Clore (1973), Study 2	Giving a ride to a hitchhiker	1.42 (0.20/0.01)	1.20/1.64	3/3/2/2	0.73	0.91	0.66	0.90	1.39
5. Borofsky, Stollak & Messe (1971)	Stopping a brutal fight between 2 subjects	1.23 (0.48/0.10)	0.57/1.89	1/3/2/2	0.85	0.49	0.52	0.76	0.98
6. J.A. Piliavin & I.M. Piliavin (1972)	Helping a man who fell in the subway	1.03 (0.08/0.00)	0.88/1.18	3/3/2/2	0.52	-0.03	0.34	0.23	0.00
7. Kleinke, MacIntire & Riddle (1978), Study 1	Mailing a letter for a woman in a shopping mall.	0.86 (0.85/0.57)	0.41/1.31	3/2/2/1	-0.09	-0.28	-0.03	-0.13	-0.13
8. Solomon & Herman (1977)	Picking up fallen groceries for a woman at her car.	0.79 (0.53/0.24)	0.26/1.31	3/2/1/2	0.03	-0.11	0.13	-0.07	0.26
9. Smith, Wheeler & Diener (1975)	Volunteering to spend time with retarded children	-0.70 (0.07/0.22)	-0.91/-0.50	2/1/2/1	-0.29	-0.63	-0.28	-0.55	-1.29
10. Austin (1979), Study 3	Stopping someone from stealing a student's belongings in a classroom building	-0.71 (0.47/0.74)	-0.92/-0.49	2/2/1/1	0.46	0.09	-0.12	0.04	0.64

Note. Studies can be located in this table by referring to the Appendix, where the studies' sequence numbers appear. Studies with similar or identical behavior descriptions may differ on study variables because they differ on features that are not conveyed in the summary descriptions. CI = confidence interval.

[a]Summary of description given to subjects who rated behaviors. [b]Effect sizes are positive for differences in the male direction and negative for differences in the female direction; values in parentheses are the proportion of men who helped divided by the proportion of women who helped. [c]The first variable setting (1 = laboratory, 2 = campus, and 3 = off-campus); the second variable is surveillance (1 = no surveillance, 2 = unclear, and 3 = surveillance); the third variable is the availability of other helpers (1 = not available or unclear, 2 = available); and the fourth variable is type of appeal (1 = direct request, 2 = presentation of need). [d]Values are positive for differences expected to be associated with greater helping by men (greater male estimates of competence, of comfort, and of own likelihood of helping; greater female estimate of danger to self). [e]Values are positive when questionnaire respondents believed that men were more helpful than women.

REFERENCES

Aikenhead, G. S. (1973). *The interpretation of student performance on evaluative tests*. Saskatoon, Canada: Saskatchewan University. (ERIC Document Reproduction Service No. ED 013 371)

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423-437.

Bozarth, H. D., & Roberts, R. R., Jr. (1972). Signifying significant significance. *American Psychologist, 27*, 774-775.

Campbell, D. T. (1969). Definitional versus multiple operationalism. *Et Al., 2*, 14-17.

Cooper, H. M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology, 37*, 131-146.

Cooper, H. M. (1984). *The integrative research review: A systematic approach*. Beverly Hills, CA: Sage.

Crain, R. L., & Mahard, R. E. (1983). The effect of research methodology on desegregation-achievement studies: A meta-analysis. *American Journal of Sociology, 88*, 839-854.

Eagly, A. H., & Carli, L. L. (1981). Sex of researchers and sex-typed communications as determinants of sex differences in influenceability: A meta-analysis of social influence studies. *Psychological Bulletin, 90*, 1-20.

Eagly, A. H., & Crowley, M. (1986). Gender and helping behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin, 100*, 283-308.

Giaconia, R. M., & Hedges, L. V. (1982). Identifying features of effective open education. *Review of Educational Research, 52*, 579-602.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1-20.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics, 9*, 61-85.

Hedges, L. V. (1987). The meta-analysis of test validity studies: Some new approaches. In H. Braun and H. Wainer (Eds.), *Test validity for the 1990s and beyond*. New York: Erlbaum.

Hedges, L. V., & Olkin, I. (1980). Vote counting methods in research synthesis. *Psychological Bulletin, 88*, 359-369.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Hipscher, W. L. (1961). Study of high school physics achievement. *The Science Teacher, 28* (6), 36-37.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating findings across research*. Beverly Hills, CA: Sage.

Light, R. J. (Ed.). (1983). *Evaluation studies review annual* (Vol. 8). Beverly Hills, CA: Sage.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Linn, M. C., & Peterson, A. C. (1985). Emergence and characterization of sex differences in spatial ability. *Child Development, 56*, 1479-1498.

Madow, W. G., Nisselson, H., & Olkin, I. (1983). *Incomplete data in sample surveys: Vol. 1. Report and case studies*. New York: Academic Press.

Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology, 64*, 553-557.

Rosenthal, R. (Ed.). (1980). *New directions for methodology of social and behavioral science: Quantitative assessment of research domains* (No. 5). San Francisco: Jossey-Bass.

Rosenthal, R. (1984). *Meta-analytic procedures for social research.* Beverly Hills, CA: Sage.

Shymansky, J. A., Hedges, L. V., Woodworth, G. G., & Berg, C. (1986, March 29). *A study of uncertainties in the meta-analysis of research on the effectiveness of 'new' science curricula.* Paper set presented at the annual meeting of the Association for Research in Science Teaching, San Francisco, CA.

Smith, M. L. (1980). Publication bias in meta-analysis. *Evaluation in education: An international review series, 4,* 22-24.

Thomas, J. R., & French, K. E. (1985). Gender differences across age in motor performance: A meta-analysis. *Psychological Bulletin, 98,* 260-282.

Tukey, J. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Vanek, E. A. P. (1974). A comparative study of selected science teaching materials (ESS) and a textbook approach on classificatory skills, science achievement, and attitudes (Doctoral dissertation, The University of Rochester, 1974). *Dissertation Abstracts International, 32,* 3583A. (University Microfilms No. 72-3, 721)

Webb, E., Campbell, D., Schwartz, R. Sechrest, L. & Grove, J. (1981). *Unobtrusive measures: Nonreactive research in the social sciences.* Boston: Houghton Mifflin.

White, C. (1973). *Sources of information in the social sciences.* Chicago: American Library Association.

Wideen, M. F. A product evaluation of Science—A Process Approach (Doctoral dissertation, University of Colorado, 1971). *Dissertation Abstracts International, 32,* 3583A. (University Microfilms No. 72-3, 721)

Yeaton, W. H., & Wortman, P. M. (Eds.). (1984). *New directions in program evaluation: Issues in data synthesis* (No. 24). San Francisco: Jossey-Bass.

APPENDIX I

META-ANALYSIS CODE BOOK

Each unit of observation (one effect size in one study) comprises two lines of data in the raw data file, each study generally yields several lines of data. The data items on each of the lines are as follows:

Data Item	Description
1. Study ID	Identification number for the study.
2. Subgroup	Code letter identifying a subgroup of pupils.
3. Scale	Code number identifying the criterion measure.
4. Pupil Assignment	Method of assignment of pupils to treatments. (1) Random (2) Matched (3) Intact Classes (4) Self-selection
5. Teacher Assignment	Method of assignment of teachers to treatments. (1) Random (2) Non-random (3) Self-selected (4) Crossed (5) Matched
6. Criterion Measure	Type of criterion measure (1) Cognitive - low (2) Cognitive - high (3) Cognitive - mixed (4) Problem solving (5) Affective - subject (6) Affective - science (7) Affective - procedure/methodology (8) Values (9) Process skills, techniques (10) Methods of science (11) Psychomotor (12) Critical thinking (13) Creativity (14) Decision making (15) Logical thinking, Piagetian (16) Spatial relations, Piagetian (17) Self-Concept (18) Classroom behaviors (19) Reading (20) Mathematics (21) Social studies (22) Communication skills
7. Criterion Cluster	Broader classification of criterion measure. (1) Achievement (2) Perceptions (3) Process Skills (4) Analytic Skills (5) Other
8. Curriculum	New science curricula Elementary (1) ESS (2) SCIS, SCIIS, SCIS II (3) S-APA (4) OBIS (5) ESLI (6) ESSENCE (7) COPES (8) MAPS (9) USMES (10) MINNEMAST (11) IS (12) SCIL (15) ESTPSI (16) FHESP

(codebook, continued)

Junior High

(28) HSP	(29) TSM
(30) ISIS	(31) ISCS
(33) IPS	(34) ESCP
(35) IME	(36) CE/EE
(37) MSP	

Secondary

(50) BSCS SPECIAL	(51) BSCS YELLOW
(52) BSCS BLUE	(53) BSCS GREEN
(54) BSCS ADVANCED	(55) CHEM STUDY
(56) CBA	(57) PSSC
(58) HPP	(59) CE/EE
(60) PSNS	(61) IAC

9. Grade Level — (1) K-3 (2) 4-6 (3) 7-9
(4) 10-12 (5) Post Secondary

10. Inservice — Was inservice training in new curriculum provided?
(1) Yes (2) No

11. IQ — Average IQ of students
(1) below 95 IQ (2) 95-105 (3) above 105

12. Length — Length of study
(1) under 14 weeks (2) 14 to 28 weeks
(3) over 28 weeks

13. Preservice — Was preservice training provided?
(1) Yes (2) No

14. School Size — Number of pupils in school.
(1) < 50 (2) 50-199 (3) 200-499 (4) 500-999
(5) 1000-1999 (6) > 2000

15. School Type — (1) Rural (2) Suburban (3) Urban

16. SES — Average socioeconomic status of pupils in treatment and control groups. (1) Low (2) Medium (3) High

17. SEX — Sex ratio of treatment and control groups.
(1) over 75% male (2) over 75% female
(3) at least 25% of either sex

18. Test type — (1) Standardized (2) Ad hoc written test
(3) Classroom test (4) Observation
(5) Structured interview

19. Test Content — (1) Life science (2) Physical science
(3) General science (4) Earth science (5) Biology
(6) Chemistry (7) Physics

(codebook, continued)

20.	Teacher Background	Educational background of treatment group teacher. (1) < Bachelors (2) Bachelors (3) Bachelors +35 (4) Masters (5) Masters +15 (6) Masters +30 (7) Doctorate
21.	Teacher Experience	Years of science teaching experience.
22.	CFLAG	Was the analysis covariate adjusted? (1) Yes, impossible to deduce unadjusted Std Dev

(The following items are on the second line of data for each observation.)

23.	Treatment	Mean
24.		Standard Deviation
25.		Sample Size
26.	Control	Mean
27.		Standard Deviation
28.		Sample Size
29.	AN(C)OVA	Sum of Squares for Curricula
30.		Degrees of Freedom for Curricula (should be 1)
31		Total Sum of Squares
32.		Total Degrees of Freedom
33.		F statistic for Curricula
34.	Sign	Sign of the effect. (1) Treatment better than control (-1) Control better than treatment

APPENDIX II

POOLING CORRELATED EFFECT SIZES

Notation:

Say that in a particular study subscales A, B and C are to be pooled into one effect size. The individual effect size estimates and their standard errors are d_A s_A, d_B s_B and d_C s_C. The pooled effect size estimate is the average,

$$d_{pooled} = (d_A + d_B + d_C)/3.$$

It can be proved that regardless of the correlations among the subscales, the standard error of this pooled estimate is less than or equal to the average of their three standard errors,

$$s_d \leq (s_A + s_B + s_C)/3,$$

consequently, the average standard error is a conservative estimate of the standard error of the pooled effect size.

APPENDIX III

COMPUTER SAS PROGRAMS

This appendix contains computer programs for carrying out the statistical analyses described in this handbook via the SAS (Statistical Analysis System) package of programs. SAS is among the most widely available statistical computer packages for mainframe computations. Microcomputer versions of SAS are also available. In this appendix, we present complete programs to carry out an entire analysis. We identify the functions of segments of code that accomplish particular purposes. These segments are designed to illustrate how to use SAS to carry out particularly tricky or unfamiliar operations. They must be combined and modified to carry out the particular analysis desired in any meta-analysis.

Program I. Data Entry and Effect Size Computation.

```
// EXEC SAS, OPTIONS='NOCENTER NODATE'
//METADISK DD DSN=META.SAS,UNIT=DISK,DISP=(NEW,CATLG),
//   SPACE=(TRK, (50,10))
//SYSIN DD*

DATA METADISK.META ;                    *See CODEBOOK for data description;
  INPUT (ID SUBGRP SCALE ASSIGNSS ASSIGNTE CRITERIA CLUST CURRIC
    GRADE INSERV IQ LENGTH PRESREV SCHSIZE SCHTYPE SES SEX
          TESTTYPE
    TESTCONT TEABCKGD EXPERT CFLAG TMEAN TSTD TN CMEAN CSTD CN
          SSTREAT
    DFTREAT SSTOT DFTOT FTREAT SGN)
    (+5 3.0 +1 $1. 3*2.0 3.0 2.0 3.0 14*2.0/
     +7 7.0 6.0 4.0 7.0 6.0 4.0 10.0 2.0 12.0 5.0 7.0 2.0) ;

IF TN=. OR CN=. THEN DO ;               *Assume equal samples if not stated ;
   TN=(DFTOT+1)/2 ;
   CN=TN ;
END ;

IF DFTOT=. THEN DFTOT=TN+CN-1 ;         *Compute total degrees of freedom ;

IF SSTREAT NE . AND SSTOT NE . THEN
FTREAT=(SSTREAT/SSTOT-SSTREAT))*(DFTOT-1) ; *Compute F                    ;

IF TMEAN NE . AND CMEAN NE . AND CSTD NE .
  THEN G = (TMEAN - CMEAN)/CSTD ;  *Hedges & Olkin 5.A.2(3)               ;
  ELSE G = SGN*SQRT(TN+CN)*FTREAT/(TN*CN)) ;*See footnote 1               ;

IF CSTD = . OR CSTD = TSTD
     THEN DF = DFTOT - 1 ;  *Compute degrees of freedom for effect size  ;
     ELSE DF = CN - 1 ;     *  df = m in Hedges & Olkin 5.A.2(7).        ;

J = 1-3/(4*DF - 1) ;                    * D is the unbiased estimate of the effect ;
D = G*J ;                               *   size using Hedges & Olkin 5.A.2(9,10).  ;

VARD = J*J*DF*(TN+CN)/((DF-2)*TN*CN)+D*D*(J*J*DF/(DF-2)-1) ;
STD = SQRT(VARD) ;              *Standard Deviation of D from H & 0 5.E(36).  ;

CARDS ;       *Beginning of data (see CODEBOOK for variable description)    ;

     ***** DATA GOES HERE *****


;;;;
```

Program II. Store Value Labels in SAS

```
// EXEC SAS5, OPTIONS='NOCENTER NODATE'
//SASLIB DD DSN=SFOFMT.SAS,UNIT=DISK,DISP=(NEW,CATLG) ,
//   SPACE=(TRK,(4,2,2))
//SYSIN DD *
PROC FORMAT LIBRARY=SASLIB  ;
  VALUE TTYFMT  1='STANDARDIZED TEST'
                2='AD HOC WRITTEN TEST'
                3='CLASSROOM TEST'
                4='OBSERVATION'
                5='INTERVIEW';
  VALUE CNTFMT  1='LIFE SCIENCE'
                2='PHYSICAL SCIENCE'
                3='GENERAL SCIENCE'
                4='EARTH SCIENCE'
                5='BIOLOGY'
                6='CHEMISTRY'
                7='PHYSICS';
  VALUE CRCFMT  01='ESS'
                02='SCIS'
                03='S-APA'
                04='OBIS'
                05='ESLI'
                06='ESSENCE'
                07='COPES'
                08='MAPS'
                09='USMES'
                10='MINNEMEAST'
                11='IS'
                12='SCIL'
                15='ESTPSI'
                16='FHESP'
                28='HSP'
                29='TSM'
                30='ISIS'
                31='ISCS'
                33='IPS'
                34='ESCP'
                35='IME'
                36='CE/EE'
                37='MSP'
                50='BSCS-S'
                51='BSCS-Y'
                52='BSCS-B'
                53='BSCS-G'
                54='BSCS-A'
                55='CHEM STUDY'
                56='CBA'
                57='PSSC'
                58='HPP'
                59='CE/EE'
                60='PSNS'
                61='IAC'
```

(value labels)

```
VALUE GRDFMT    1='K-3'
                2='4-6'
                3='7-9'
                4='10-12'
                5='POST SECONDARY';
VALUE IQ_FMT    1='LOW (BELOW 95)'
                2='AVERAGE (95-105)'
                3='HIGH (ABOVE 105)';
VALUE SESFMT    1='LOW'
                2='MIDDLE'
                3='HIGH';
VALUE CRTFMT    1='COGNITIVE - LOW'
                2='COGNITIVE - HIGH'
                3='COGNITIVE - MIXED'
                4='PROBLEM SOLVING'
                5='AFFECTIVE - SUBJECT'
                6='AFFECTIVE - SCIENCE'
                7='AFFECTIVE PROCEDURE/METHODOLOGY'
                8='VALUES'
                9='PROCESS SKILLS, TECHNIQUES'
                10='METHODS OF SCIENCE'
                11='PSYCHOMOTOR'
                12='CRITICAL THINKING'
                13='CREATIVITY'
                14='DECISION MAKING'
                15='LOGICAL THINKING PIAGETIAN'
                16='SPATIAL RELATIONS PIAGETIAN'
                17='SELF-CONCEPT'
                18='CLASSROOM BEHAVIORS'
                19='READING'
                20='MATHEMATICS'
                21='SOCIAL STUDIES'
                22='COMMUNICATION SKILLS';
VALUE CC_FMT    1='ACHIEVEMENT CLUSTER'
                2='PERCEPTIONS CLUSTER'
                3='PROCESS SKILLS'
                4='ANALYTIC SKILLS'
                5='OTHER PERFORMANCE AREAS'
                6='RELATED SKILLS';
VALUE SEXFMT    1='OVER 75% MALE'
                2='OVER 75% FEMALE'
                3='AT LEAST 25% MALES AND FEMALES';
VALUE LNGFMT    1='UNDER 14 WEEKS'
                2='14 TO 28 WEEKS'
                3='OVER 28 WEEKS';
```

Program III. Meta Analysis and Diagnostic Plots

```
// EXEC SAS,OPTIONS='PAGESIZE=330 NOCENTER NODATE'
//METADISK DD DSN=USER.C5001420.META.SAS,UNIT=DISK,DISP=SHR
//SASLIB DD DSN=USER.C5001420.METAFMT.SAS,UNIT=DISK,DISP=SHR
//SYSIN DD *

*-------------------------------------------------------------------------------;
*    G.METANL          2/10/87          GENERIC META ANALYSIS PROGRAM;
*-------------------------------------------------------------------------------;

MACRO FACTORS
  CURRIC SEX IQ
%

DATA METAFMTD ;
SET METADISK.META ;
  IF CFLAG THEN DO;  D=D*.7; STD=STD*.7; END;
  GRAND=1;

  FORMAT TESTTYPE TTYFMT.
         TESTCONT CNTFMT.
         CURRIC   CRCFMT.
         GRADE    GRDFMT.
         SEX      SEXFMT.
         IQ       IQ_FMT.
         SES      SESFMT.
         CRITERIA CRTFMT.
         CLUST    CC_FMT.
         LENGTH   LNGFMT.;

PROC SORT DATA=METAFMTDD ;
  BY CLUST  ID  SUBGRP  ;

PROC MEANS NOPRINT DATA=METAFMTD ; * AGGREGATE CORRELATED ;
  BY CLUST ID SUBGRP  ;                * EFFECT SIZES.          ;
  ID FACTORS  ;
  VAR D STD  ;
  OUTPUT OUT=AGGREGAT
         MEAN= D STD  ;

DATA AGGREGAT  ;
  SET AGGREGAT  ;
  STDSQR=STD*STD  ;                    * CREATE WEIGHTS       ;
  WT=1/STDSQR  ;

PROC SORT DATA=AGGREGAT  ;
  BY CLUST FACTORS D ;
```

(program III, continued)

```
PROC MEANS NOPRINT DATA=AGGREGAT  ;
  BY CLUST FACTORS ;
  VAR D WT  ;
  WEIGHT WT  ;
  OUTPUT OUT=SUMMARY
         MEAN=DBAR WTBAR
         CSS=Q
         SUMWGT=SUMWGT
         N=N  ;

PROC MEANS NOPRINT DATA=AGGREGAT  ;
  BY CLUST FACTORS  ;
  VAR D STDSQR  ;
* UNWEIGHTED  ;
  OUTPUT OUT = USUMMARY
    VAR  = UNWVAR
    MEAN = JUNK  SDSQBR  ;

DATA SUMMARY ;
  MERGE SUMMARY USUMMARY  ;
  PCHI = .  ;
  IF N GT 1 THEN DO  ;
    PCHI=1-PROBCHI(Q,N-1)  ;
  END  ;
  SIG_DELT = SQRT (MAX((UNWVAR-SDSQBR) ,0)) ;
  TOT_SE = SQRT((WTBAR*(SIG_DELT**2) + 1)/SUMWGT) ;
  SAMP_SE = 1/SQRT(SUMWGT)  ;

PROC PRINT  ;
  BY CLUST  ;
  VAR FACTORS DBAR TOT_SE SAMP_SE N Q PCHI  ;

DATA PLOTDATA  ;
  MERGE AGGREGAT SUMMARY(KEEP=CLUST FACTORS DBAR);
  BY CLUST FACTORS  ;
  N=_N_
  EFFECT=D;
  UPPER=D+3*STD;
  LOWER=D-3*STD;
  RESID = WT*(D-DBAR)**2  ;

PROC TIMEPLOT  ;
  BY CLUST  ;
  PLOT LOWER='L' EFFECT='*'  UPPER='U'  DBAR='|' / HILOC OVERLAY;
  ID SUBGRP FACTORS RESID  ;

PROC GLM DATA=AGGREGAT ;
  TITLE 'ANALYSIS OF HETEROGENEITY BY WEIGHTED ANOVA'  ;
  BY CLUST ;
  CLASS FACTORS  ;
  WEIGHT WT  ;
  MODEL D = FACTORS SEX*IQ  ;
```

APPENDIX IV

ACHIEVEMENT EFFECT SIZES AND HETEROGENEITY PLOTS FROM SHYMANSKY, HEDGES, AND WOODWORTH

Achievement Effect Size Estimates

Study ID : Subgroup : Pupil Assignment : Teacher Assignment : Grade : Inservice : IQ : Study Length : Preservice : School Size : School Type : Socioeconomic Status : Sex : Test Type : Test Content : Covariate

See Appendix 1 for coding of variables

																	Effect Size	Standard Deviation	notes
ESS																			
42	A	3	5	2	1	2	2	.	.	.	2	3	1	3	.		0.41072	0.161134	grade 4
42	B	3	5	2	1	2	2	.	.	.	2	3	1	3	.		0.13609	0.148538	grade 5
42	C	3	5	2	1	2	2	.	.	.	2	3	1	3	.		-0.45124	0.163649	grade 6
102	A	1	4	2	1	2	1	.	4	1	2	1	1	3	.		0.04093	0.191164	
SCIS																			
2	A	3	1	1	.	2	2	.	.	2	2	3	5	2	.		1.12835	0.148074	SCIS test
78	A	3	3	2	.	1	2	.	.	3	1	3	2	3	.		0.78155	0.420886	
SAPA																			
5	F	3	2	2	.	2	3	.	.	2	2	3	1	3	0		-0.70392	0.176314	
15	A	3	2	2	1	2	3	.	.	2	2	1	1	3	.		0.18290	0.130201	
15	B	3	2	3	1	2	3	.	.	2	2	1	1	3	.		-0.01988	0.127055	
84	A	.	.	2	1	2	1	2	.	2	2	3	1	3	.		-0.01738	0.260621	
85	B	3	2	2	1	2	3	2	.	.	2	3	1	3	.		0.00640	0.204681	
96	A	3	3	2	2	2	3	2	.	.	2	1	1	3	2		-0.12007	0.088654	
103	A	3	2	2	1	3	3	.	.	2	2	1	1	3	.		0.35791	0.380184	hi-IQ boys
103	B	3	2	2	1	1	3	.	.	2	2	1	1	3	.		0.05436	0.373062	lo-IQ boys
103	C	3	2	2	1	3	3	.	.	2	2	2	1	3	.		-0.49585	0.386766	hi-IQ girls
103	D	3	2	2	1	1	3	.	.	2	2	2	1	3	.		1.56206	0.493599	lo-IQ girls
106	A	3	3	2	.	3	2	.	.	2	2	3	1	2	.		0.27856	0.085871	
MINNEMAST																			
12	A	3	2	1	.	2	1	.	.	2	2	3	3	3	0		1.72503	0.165242	K, passive ctrl
ESTPSI																			
93	A	3	2	2	.	2	2	.	.	.	2	3	2	3	.		0.28837	0.157448	grade 5
93	B	3	2	2	.	2	2	.	.	.	2	3	2	3	.		0.35702	0.163749	grade 6
93	C	3	2	3	.	2	2	.	.	.	2	3	2	3	.		0.31961	0.200051	grade 7
93	D	3	2	3	.	2	2	.	.	.	2	3	2	3	.		0.02743	0.194970	grade 8
FHESP																			
55	A	3	2	2	1	2	3	1	.	1	2	3	1	3	.		0.07195	0.121300	grade 6?
55	B	3	2	3	1	2	3	1	.	1	2	3	1	3	.		0.13537	0.151229	grade 7?
55	C	3	2	3	1	2	3	1	.	1	2	3	1	3	.		1.00996	0.216737	grade 8?

(achievement effects, continued)

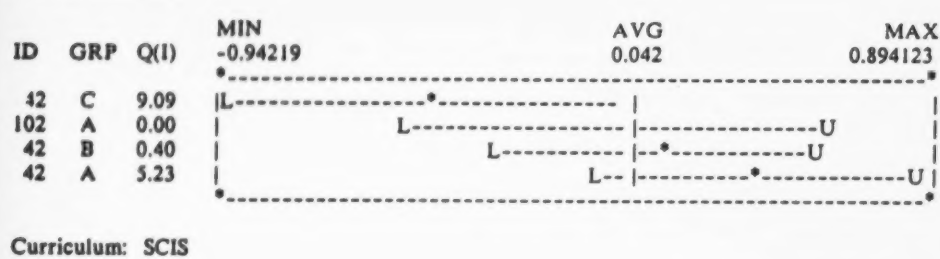
Study ID	Subgroup	Pupil Assignment	Teacher Assignment	Grade	Inservice	IQ	Study Length	Preservice	School Size	School Type	Socioeconomic Status	Sex	Test Type	Test Content	Covariate	Effect Size	Standard Deviation	notes
IPS																		
59	B	3	2	3	1	2	3	1	.	2	2	3	1	2	.	0.06793	0.144511	
90	A	2	2	3	.	2	3	.	.	3	2	1	1	2	.	0.55793	0.163561	
ESCP																		
53	A	3	2	3	2	2	2	.	.	2	2	3	1	2	.	0.74036	0.119597	
92	A	3	5	4	.	2	3	.	1	2	2	3	2	4	.	-0.04322	0.073760	
IME																		
59	A	3	2	3	1	2	3	1	.	2	2	3	1	2	.	0.19937	0.128919	
MSP																		
89	A	3	2	3	.	2	3	.	.	2	2	3	1	2	.	0.49103	0.059489	
BSCS SPECIAL																		
105	A	3	3	4	.	1	2	.	.	.	2	3	1	5	.	0.68906	0.163130	bscs test
105	B	3	3	4	.	2	2	.	.	.	2	3	1	5	.	0.14534	0.179609	bscs test
BSCS YELLOW																		
9	A	3	2	4	1	2	3	.	.	3	2	1	2	5	1	0.67660	0.069420	israel-city
9	B	3	2	4	1	2	3	.	.	2	2	1	2	5	1	0.93613	0.092565	israel-kibbutz
9	C	3	2	4	1	2	3	.	.	1	2	1	2	5	1	0.03495	0.090629	israel-ag schl
29	A	3	2	4	.	3	2	.	.	.	2	1	1	5	.	0.44184	0.096960	
37	A	3	2	4	.	3	2	.	.	2	2	3	3	5	.	0.28069	0.140289	zoology classes
37	B	3	2	4	.	3	2	.	.	2	2	3	3	5	.	0.52985	0.142355	botany classes
37	C	3	2	4	.	3	2	.	.	2	2	3	3	5	.	0.38884	0.141362	biology classes
43	A	1	2	4	.	3	3	.	.	1	2	3	1	5	.	0.56643	0.165552	
51	A	3	2	4	.	2	3	.	.	.	2	3	1	5	.	0.21218	0.061099	
63	A	3	2	4	.	2	1	.	.	2	2	3	2	5	.	0.71001	0.149950	
65	B	3	2	4	.	2	3	.	6	2	3	3	1	5	.	0.63274	0.186029	passive control
82	A	3	2	4	1	2	3	1	4	1	2	1	1	5	.	0.79051	0.475224	bscs inquiry
82	B	3	2	4	1	2	3	1	4	1	2	2	1	5	.	0.44143	0.340299	bscs inquiry
82	C	3	2	4	2	2	3	1	4	1	2	1	1	5	.	0.02289	0.392769	bscs traditional
82	D	3	2	4	2	2	3	1	4	1	2	2	1	5	.	-0.81212	0.353206	bscs traditional
BSCS BLUE																		
65	A	3	2	4	.	3	3	.	6	2	3	3	1	5	.	1.01498	0.268030	passive control

(achievement effects, continued)

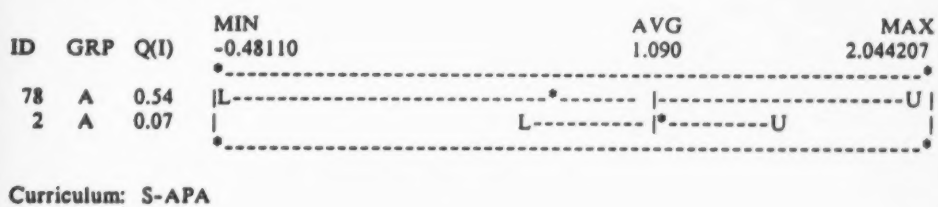
Study ID	Subgroup	Pupil Assignment	Teacher Assignment	Grade	Inservice	IQ	Study Length	Preservice	School Size	School Type	Socioeconomic Status	Sex	Test Type	Test Content	Covariate	Effect Size	Standard Deviation	notes
BSCS GREEN																		
76	A	3	3	4	.	2	2	.	.	2	2	3	1	5	.	0.01219	0.086954	
BSCS ADVANCED																		
77	A	3	3	4	1	3	2	.	.	2	2	1	1	5	.	0.12980	0.146706	
77	B	3	3	4	1	1	2	.	.	2	2	1	1	5	.	0.04786	0.266529	
CHEM STUDY																		
14	A	4	2	4	.	3	2	.	.	.	2	1	2	6	.	-0.40871	0.121585	
39	A	2	2	4	.	3	2	.	.	2	2	1	1	6	.	0.09753	0.168244	
40	A	3	2	4	1	3	2	.	.	.	2	1	1	6	1	-0.24486	0.105943	
40	B	3	2	4	1	2	2	.	.	.	2	1	2	6	1	-0.24595	0.151661	
40	C	3	2	4	1	1	2	.	.	.	2	1	2	6	1	-0.19057	0.329524	
58	A	3	2	4	.	3	3	.	6	2	3	3	1	6	.	-0.29619	0.077485	
66	A	3	2	4	2	3	1	2	.	2	2	1	1	6	.	0.30718	0.136083	short study
70	A	3	3	4	.	3	2	.	.	2	3	1	2	6	.	0.14979	0.156307	
70	B	3	3	4	.	2	2	.	.	2	3	1	2	6	.	-0.10863	0.170477	
70	C	3	3	4	.	1	2	.	.	2	3	1	2	6	.	-0.02463	0.151612	
101	A	3	2	4	.	3	2	.	.	.	2	3	2	6	.	0.31476	0.059489	
CBA																		
17	A	3	2	4	.	3	3	.	.	.	2	3	2	6	.	0.89192	0.107136	
32	A	3	1	4	.	3	2	.	.	.	2	1	1	6	.	0.48603	0.294465	
81	A	3	2	4	1	3	3	1	.	2	2	3	1	6	.	0.49025	0.291222	
101	B	3	2	4	.	3	2	.	.	.	2	3	1	6	.	0.28714	0.091950	
PSSC																		
8	A	3	2	4	.	3	3	.	.	2	2	1	1	7	.	-0.72608	0.103870	
22	A	3	2	4	.	3	2	.	.	.	2	1	1	7	.	0.26598	0.074976	
36	A	3	2	4	1	3	2	.	.	2	2	1	2	7	.	0.54055	0.079886	
45	A	3	2	4	1	3	2	.	.	2	2	1	1	7	.	0.80276	0.180314	
57	A	4	2	4	.	3	3	.	.	3	2	1	2	7	.	1.09673	0.105650	
69	A	3	2	5	.	3	2	.	.	.	2	1	1	7	.	0.15886	0.452745	
71	A	3	2	5	2	3	3	2	6	2	2	1	3	7	1	0.19737	0.138334	
87	A	3	2	4	1	3	3	1	.	2	2	1	2	7	.	0.25700	0.105525	
91	A	3	3	4	.	3	2	.	.	.	2	3	2	7	.	-0.25541	0.223548	
94	A	3	2	4	.	3	1	.	.	2	2	1	1	7	.	1.19626	0.205658	
94	B	3	2	4	.	2	1	.	.	2	2	1	1	7	.	0.51193	0.121484	
94	C	3	2	4	.	1	1	.	.	2	2	1	1	7	.	0.36114	0.168377	

Heterogeneity Plots: Achievement Effect Size Estimates with Three Standard Error Bars, all Curricula

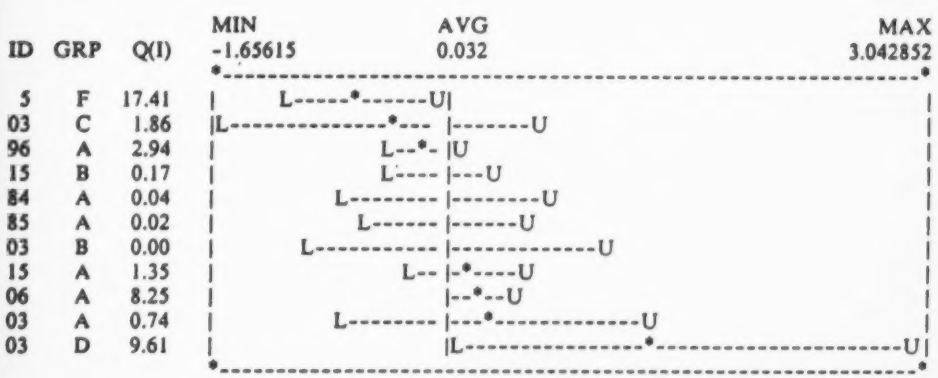
Curriculum: ESS



Curriculum: SCIS



Curriculum: S-APA

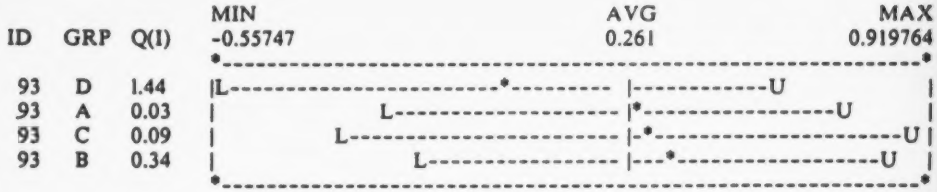


Key: $U = d_i + 3*s_i$ (Effect size + 3 sampling standard errors)
$* = d_i$ (Effect size)
$L = d_i - 3*s_i$ (Effect size - 3 sampling standard errors)
| = d (Weighted average effect size for studies of this curriculum)

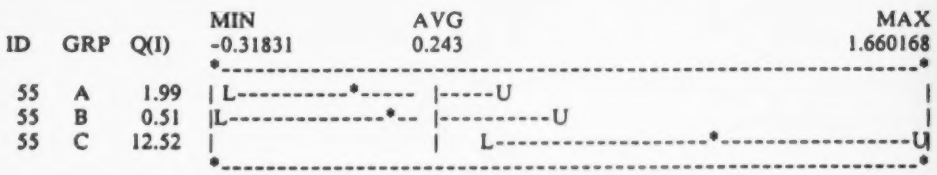
ID = Study IDentification code.
GRP = Identification code for independent subGRouPs of subjects. within studies
Q(i) = $(d_i - d)/s_i)2$ (Squared deviation of ith effect size from the weighted average)

(heterogeneity plots, continued)

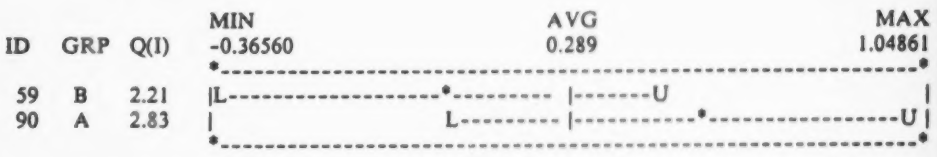
Curriculum: ESTPSI



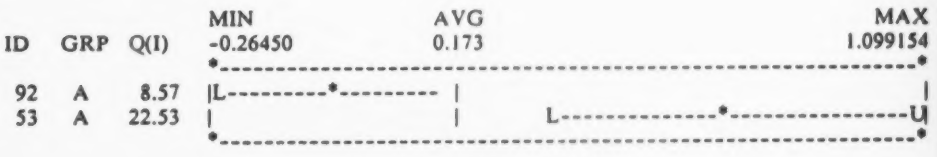
Curriculum: FHESP



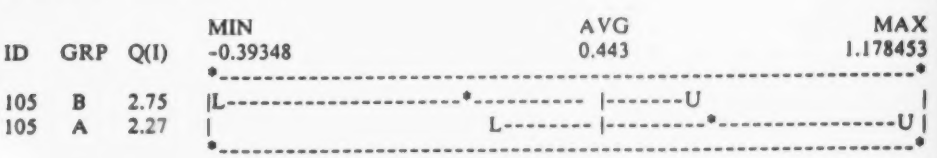
Curriculum: IPS



Curriculum: ESCP

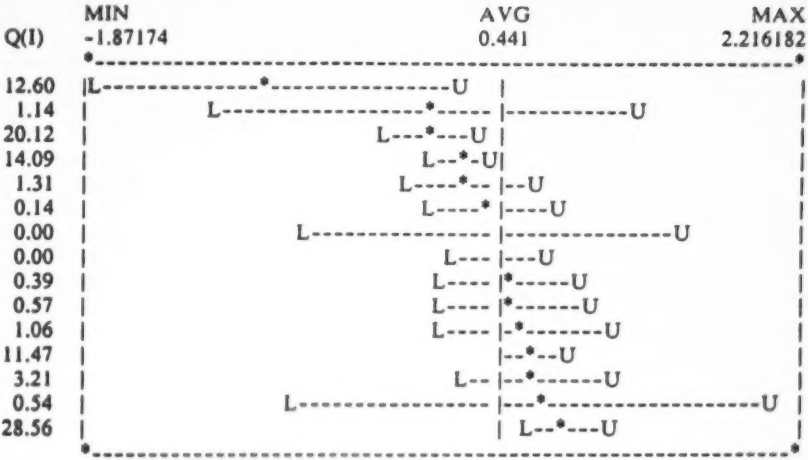


Curriculum: BSCS-S

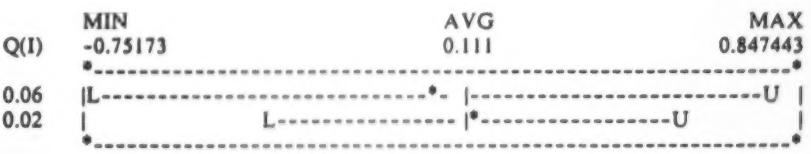


(heterogeneity plots, continued)

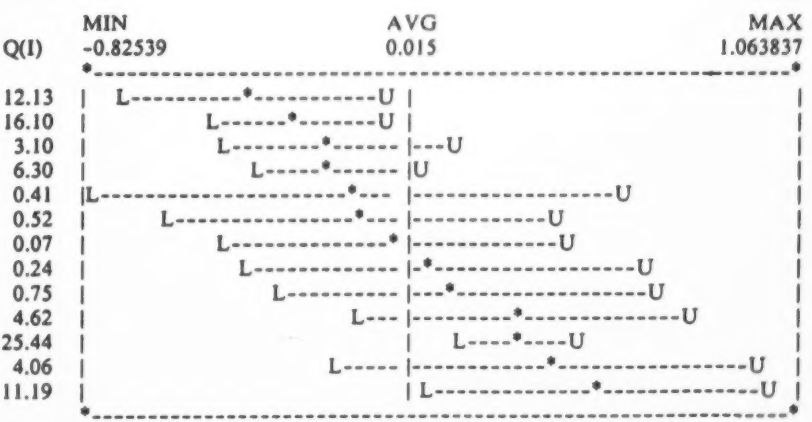
Curriculum: BSCS-Y



Curriculum: BSCS-A

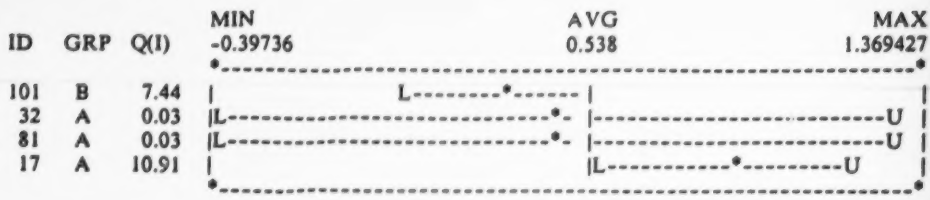


Curriculum: CHEM STUDY



(heterogeneity plots, continued)

Curriculum: CBA



Curriculum: PSSC